SEM with observed variables: parameterization and identification

Psychology 588: Covariance structure and factor models

Limitations of SEM as a causal modeling

- If an SEM model reflects the reality, the data will be consistent with the model, given that measurement errors are tolerable, all assumptions made are tenable, etc.; but the reverse is not true (e.g., see Fig. 3.9, p. 70)

- Theory- vs. data-driven

- Cause vs. effect indicators

  - simultaneous reciprocal relation (e.g., financial health and stock price of companies) --- really concurrent?

- To be cautious about goodness (mostly badness) of fit testing

Fundamental equation with observed variables

- Dependent variables $\mathbf{y}$ are modeled as linear combinations of (a subset of) $\mathbf{y}$ and $\mathbf{x}$ as

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

$$= (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta})$$

Hypothesized explanation for the covariances of $\mathbf{y}$

unexplained (but allowed to exist) in the $\mathbf{y}$ variation

- From the measurement perspective, $\mathbf{y}$ and $\mathbf{x}$ may be considered as “single-indicator latent variables” with no measurement errors

  Demographic variables are often considered so (e.g., sex, age)

Fundamental equation with observed variables

- Alternatively, $\mathbf{y}$ is modeled as linear combinations of a partitioned vector of $\mathbf{y}$, $\mathbf{x}$ and $\boldsymbol{\zeta}$ as

$$\mathbf{y} = \tilde{\mathbf{B}}\tilde{\mathbf{y}} = \left[\mathbf{B} \mid \boldsymbol{\Gamma} \mid \mathbf{I}\right]\begin{bmatrix}\mathbf{y} \\ \hline \mathbf{x} \\ \hline \boldsymbol{\zeta}\end{bmatrix} = \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

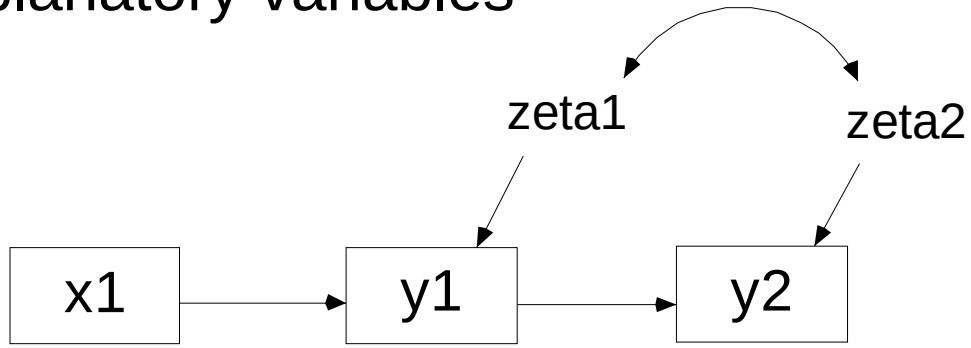
- This is essentially the approach taken by RAM (reticular action model), though RAM incorporates all latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ --- to be discussed later

Two conditions for recursive models

- $\mathbf{B}$ should be a lower triangular matrix, and so no feedback loop of causal paths

- All error terms ζ are not correlated with one another

  Correlated errors themselves don't result in a feedback loop; instead, they lead to inconsistent estimates due to errors correlated with explanatory variables



- Note: any exogenous variables (including error terms) can be set to inter-correlate while no endogenous variables are allowed so (instead, set to correlate through exogenous variables)

Implied covariance matrix

- Basic hypothesis with ideal measurement:

  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ --- population covariance matrix $\boldsymbol{\Sigma}$ is a function of free model parameters $\boldsymbol{\theta}$

- Basic hypothesis in reality:

  $\mathbf{S} = \boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}\right)$ --- sample covariance matrix $\mathbf{S}$ is a function of estimates of model parameters $\hat{\boldsymbol{\theta}}$

- Discrepancy in the data $(\boldsymbol{\Sigma}$ vs. $\mathbf{S})$ is due to sampling errors while the discrepancy in the parameters $(\boldsymbol{\theta}$ vs. $\hat{\boldsymbol{\theta}})$ is due to not only sampling errors but also any violated assumptions made for a particular way of finding the estimates (e.g., ML)

Implied covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = E\left(\begin{bmatrix}\mathbf{y}\\ \hline \mathbf{x}\end{bmatrix}[\mathbf{y}' \mid \mathbf{x}']\right) \qquad \mathbf{y} = (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\mathbf{x}+\boldsymbol{\zeta})$$

$$= \left[\begin{array}{c|c} E(\mathbf{y}\mathbf{y}') & E(\mathbf{y}\mathbf{x}') \\ \hline E(\mathbf{x}\mathbf{y}') & E(\mathbf{x}\mathbf{x}') \end{array}\right]$$

$$= \begin{bmatrix} (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'+\boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1\prime} & (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1\prime} & \boldsymbol{\Phi} \end{bmatrix}$$

Identification

Preliminaries:

- SEM is a modeling technique for covariance structures; thus structural models are written and treated in the covariance form --- indirect vs. direct fitting

- In most SEM models, we exclusively consider only the covariance matrix, not means; in addition, third and higher moments are excluded by imposing multinormality on observed variables

- Variances and covariances of observed variables are assumed to be known; thus we represent model parameters as a function of these “knowns”

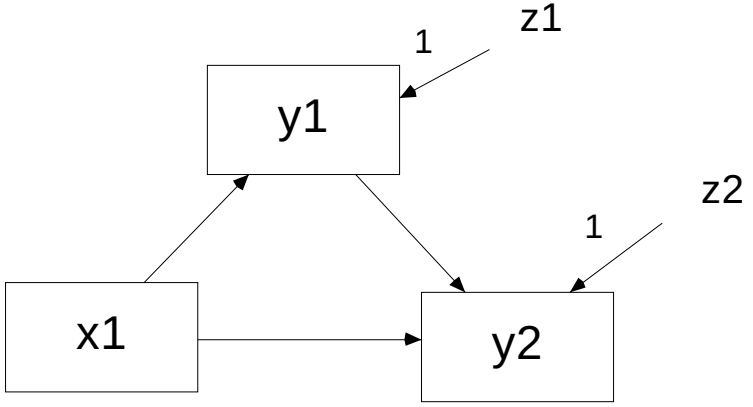
- Identification in its complete sense means that unknown model parameters are determined to take particular values (preferably unique values)

- To identify unknown parameter values, we take two steps:

  Step 1: Identifiability of model form --- mathematical reasoning of whether a given model takes a form of solvable problem

  Step 2: Estimation of the parameter values --- numerical optimization of some loss function (e.g., sum of squared residuals for the OLS regression)

- Given the degrees of freedom in the <u>covariance</u> data ("knowns"; df_1) and in the model ("unknowns"; df_2), there are 3 possibilities:
  - $df_1 = df_2$ --- just identified, data exactly reproducible ("known to be identified", "identifiable")
  - $df_1 > df_2$ --- over-identified, uniquely identifiable with additional conditions but imperfect fit
  - $df_1 < df_2$ --- under-identified, impossible to uniquely identify



Identification: t-rule

$$t \leq \frac{1}{2}(p+q)(p+q+1)$$

- t = # of distinctive free parameters in $\boldsymbol{\theta}$, $p+q$ = # of observed variables
  - equality constraints reduce t but inequality constraints don't --- inequality constraints limit the search space of the parameter, and so they may produce more accurate and reliable estimates if they are consistent with the data
- Necessary but not sufficient; only useful for ruling out unidentifiable models

- E.g., the SES model (Fig. 4.4, p. 84) meets the t-rule $(df_1 = df_2 = 15)$, but it doesn't mean it's identifiable --- we'll reconsider this model with other rules

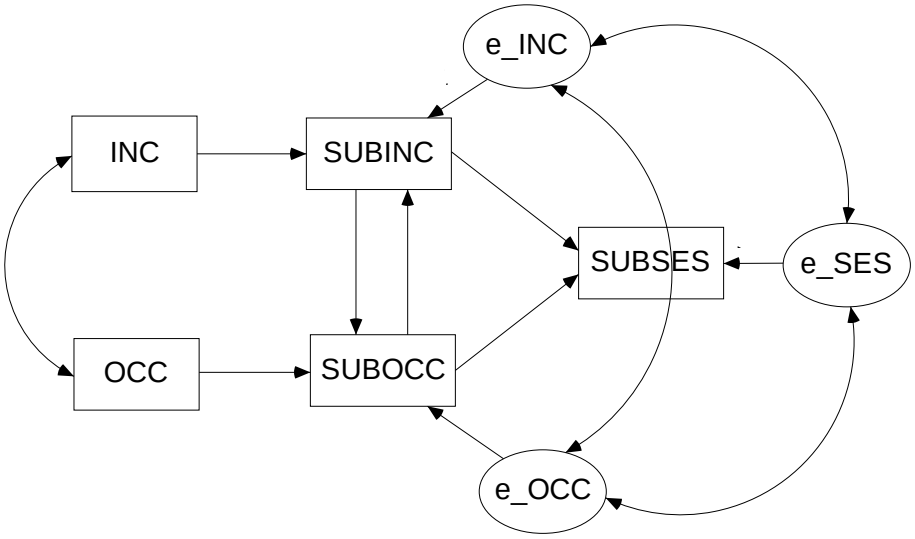
$$\mathbf{B} = \begin{bmatrix} 0 & \beta_{12} & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{bmatrix}, \quad \mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \\ 0 & 0 \end{bmatrix}, \quad \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & & \\ \psi_{21} & \psi_{22} & \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix},$$

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

x_1: INC, x_2: OCC

y_1: SUBINC, y_2: SUBOCC

y_3: SUBSES



Identification: null $\mathbf{B}$ rule

- All DVs are influenced only by IVs, not DVs; i.e., there is no mediators; and so the fundamental equation reduces to the multivariate regression model (with some regression coefficients possibly constrained to zero):

$$\mathbf{y} = \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$$

- Consequently, the implied covariance matrix reduces to:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}'_{xy} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi} & \boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}' & \boldsymbol{\Phi} \end{bmatrix}$$

- And all parameter sets (unknowns) are identified as:

$$\boldsymbol{\Phi} = \boldsymbol{\Sigma}_{xx}, \quad \boldsymbol{\Gamma} = \boldsymbol{\Sigma}'_{xy}\boldsymbol{\Sigma}_{xx}^{-1}, \quad \boldsymbol{\Psi} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}'_{xy}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

- $\mathbf{B} = \mathbf{0}$ is a sufficient condition for identification, but not necessary, and so this rule doesn't tell if models with $\mathbf{B} \neq \mathbf{0}$ are identifiable

- Furthermore, if all entries of Γ are unconstrained (i.e., the multivariate linear regression model), models with $\mathbf{B} = \mathbf{0}$ are <u>just</u> identified implying a perfect fit

  - How is the perfect fit possible if residual variances $(\psi_{11},\ldots,\psi_{pp})$ in the regression model are not zero, unless the DVs are exactly linear combinations of IVs (i.e., $\mathbf{\Psi} = \mathbf{0}$)?

- Obviously, the SES model doesn't meet the null-B rule

Identification: recursive rule

- All (fully) recursive models are identifiable: $\mathbf{B}$ is lower-triangular and $\mathbf{\Psi}$ is diagonal
- Consider a single equation for y_i in $\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}$

$$y_i = \left(\beta_{i1} y_1 + \cdots + \beta_{i,i-1} y_{i-1}\right) + \left(\gamma_{i1} x_1 + \cdots + \gamma_{iq} x_q\right) + \zeta_i$$

$$= \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\mathbf{z}_i + \zeta_i$$

where $\boldsymbol{\beta}_i'$ and $\boldsymbol{\gamma}_i'$ are non-zero elements in the i-th row of $\mathbf{B}$ and $\mathbf{\Gamma}$, and $\mathbf{z}_i$ collects corresponding y and x variables; By postmultiplying both sides by $\mathbf{z}_i'$ and taking expectation,

$$E\left(y_i \mathbf{z}_i'\right) = E\left(\left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\mathbf{z}_i \mathbf{z}_i'\right) + E\left(\zeta_i \mathbf{z}_i'\right)$$

$$E\left(y_i \mathbf{z}_i'\right) = E\left(\left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\mathbf{z}_i \mathbf{z}_i'\right) + E\left(\zeta_i \mathbf{z}_i'\right)$$

$$\boldsymbol{\sigma}'_{y_i z_i} = \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\boldsymbol{\Sigma}_{z_i z_i} + \boldsymbol{\sigma}'_{\zeta_i z_i} \quad \rightarrow \quad \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right] = \boldsymbol{\sigma}'_{y_i z_i}\boldsymbol{\Sigma}^{-1}_{z_i z_i}$$

since ζ_i (error term of y_i) is orthogonal to $\mathbf{z}_i$ (the predictors of y_i) --- a regression model form

- Likewise, by postmultiplying the y_i equation by its transpose and taking expectation (and substituting the above), we have

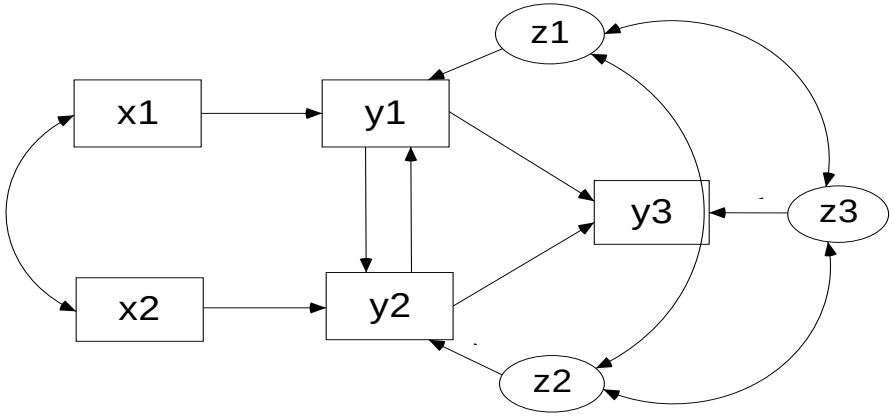
$$E\left(y_i y_i\right) = \sigma_{ii} = \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]E\left(\mathbf{z}_i \mathbf{z}_i'\right)\begin{bmatrix}\boldsymbol{\beta}_i \\ \boldsymbol{\gamma}_i\end{bmatrix} + E\left(\zeta_i \zeta_i'\right)$$

$$= \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\boldsymbol{\Sigma}_{z_i z_i}\begin{bmatrix}\boldsymbol{\beta}_i \\ \boldsymbol{\gamma}_i\end{bmatrix} + \psi_{ii} = \boldsymbol{\sigma}'_{y_i z_i}\boldsymbol{\Sigma}^{-1}_{z_i z_i}\boldsymbol{\sigma}_{y_i z_i} + \psi_{ii}$$

$$\sigma_{ii} = \boldsymbol{\sigma}'_{y_i z_i} \boldsymbol{\Sigma}^{-1}_{z_i z_i} \boldsymbol{\sigma}_{y_i z_i} + \psi_{ii} \quad \rightarrow \quad \psi_{ii} = \sigma_{ii} - \boldsymbol{\sigma}'_{y_i z_i} \boldsymbol{\Sigma}^{-1}_{z_i z_i} \boldsymbol{\sigma}_{y_i z_i}$$

- The estimation equation for $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$ is the OLS estimator (with some y as IVs), and so with proper distributional assumptions on y_i, we can do statistical testing as in regression analysis

- Recursive rule is sufficient but not necessary, and so some models with non-triangular $\mathbf{B}$ and/or non-diagonal $\boldsymbol{\Psi}$ may be identifiable; recursive rule met for the SES model?



Identification: order and rank conditions

- Useful for nonrecursive models
- No specific condition on $\mathbf{B}$ except that $(\mathbf{I} - \mathbf{B})$ is nonsingular so that $(\mathbf{I} - \mathbf{B})^{-1}$ exists
- Identification is considered one equation at a time
- No restriction in $\mathbf{\Psi}$, and so these rules would not necessarily apply to cases of restricted $\mathbf{\Psi}$ (e.g., diagonal) in that a restricted $\mathbf{\Psi}$ may help identification of otherwise unidentifiable models

Order condition

- Consider the equation for y_i similar to the one before

$$y_i = \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\mathbf{z}_i + \zeta_i$$

where $\boldsymbol{\beta}_i'$ is row i in $\mathbf{B}$ without the i-th element, $\boldsymbol{\gamma}'$ is row i in $\boldsymbol{\Gamma}$, and accordingly $\mathbf{z}_i$ collects all y and x variables except y_i

- By postmultiplying both sides by $\mathbf{x}'$ and taking expectation, we have q covariances between y_i and $\mathbf{x}$ (all IVs)

$$\boldsymbol{\sigma}_{y_i x}' = \left[\boldsymbol{\beta}_i', \boldsymbol{\gamma}_i'\right]\boldsymbol{\Sigma}_{z_i x} \quad \text{or} \quad \underset{(q\times 1)}{\boldsymbol{\sigma}_{y_i x}} = \underset{(q\times(q+p-1))}{\boldsymbol{\Sigma}_{z_i x}'} \underset{((q+p-1)\times 1)}{\begin{bmatrix}\boldsymbol{\beta}_i \\ \boldsymbol{\gamma}_i\end{bmatrix}}$$

which has q equations with $q + p - 1$ unknowns

- Consequently, if $p-1$ or more of the unknowns are excluded by zero constraints, all free parameters for the y_i equation are identifiable

- All equations' order conditions can be collectively checked with

$$\mathbf{C}=\left[\mathbf{I}-\mathbf{B},-\mathbf{\Gamma}\right]$$

If each row of $\mathbf{C}$ has $p-1$ or more zeros, it passes the order condition; satisfied for the SES model?

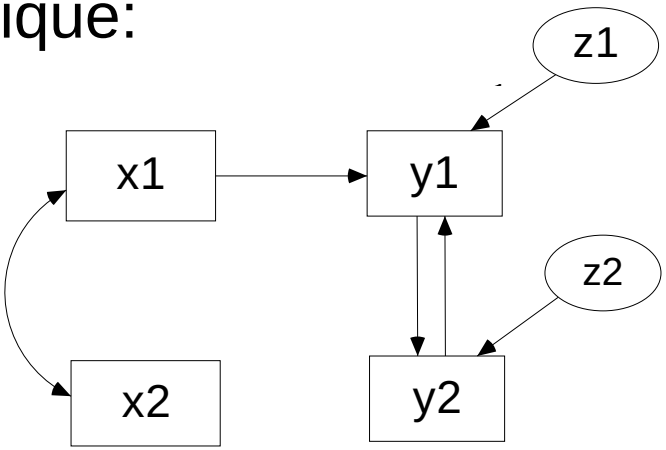
$$\mathbf{C}=\begin{bmatrix} 1 & -\beta_{12} & 0 & -\gamma_{11} & 0 \\ -\beta_{21} & 1 & 0 & 0 & -\gamma_{22} \\ -\beta_{31} & -\beta_{32} & 1 & 0 & 0 \end{bmatrix}$$

- Useful for ruling out underidentified models with unconstrained Ψ, since it's necessary but not sufficient
- A case when the order condition is met for individual equations, yet the model is not uniquely identifiable (e.g. in pp. 100-101)

$$\mathbf{C} = \begin{bmatrix} 1 & -\beta_{12} & -\gamma_{11} & 0 \\ -\beta_{21} & 1 & 0 & 0 \end{bmatrix}$$

By taking a linear combination of $\mathbf{c}_1^{*\prime} = (\mathbf{c}_1' + a\mathbf{c}_2')/(1 - a\beta_{21})$, we have an alternative solution for the first equation with the same form, implying the solution non-unique:

$$\mathbf{C}^* = \begin{bmatrix} 1 & -\beta_{12}^* & -\gamma_{11}^* & 0 \\ -\beta_{21} & 1 & 0 & 0 \end{bmatrix}$$



- For the second equation, any linear combination of the two equations can’t produce an alternative solution of the same form as the second row (i.e., the last two entries of zero), which implies that the second equation is uniquely identifiable

- The insufficiency of the order condition leads to the rank condition

Rank condition

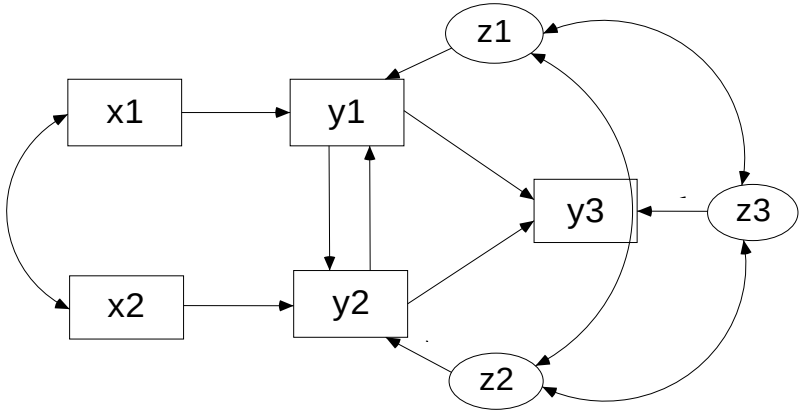
- From the $\mathbf{C}$ for the order condition, delete all columns with non-zero at row i and call the submatrix of remaining columns $\mathbf{C}_i$
- The i-th equation is identified if $\text{rank}(\mathbf{C}_i) = p - 1$
- This guarantees that the $p - 1$ equations (excluding the i-th) are all linearly independent (which are independent from equation i as well)
- The rank condition is necessary and sufficient for identification, but it doesn't take into account any restrictions on $\mathbf{\Psi}$, and so some models with restricted $\mathbf{\Psi}$ may be identifiable even if the rank condition is not met

- For the SES model in p. 84, is the rank condition met for each equation?

$$\mathbf{C} = \begin{bmatrix} 1 & -\beta_{12} & 0 & -\gamma_{11} & 0 \\ -\beta_{21} & 1 & 0 & 0 & -\gamma_{22} \\ -\beta_{31} & -\beta_{32} & 1 & 0 & 0 \end{bmatrix}$$

- How about the example of

$$\mathbf{C} = \begin{bmatrix} 1 & -\beta_{12} & -\gamma_{11} & 0 \\ -\beta_{21} & 1 & 0 & 0 \end{bmatrix}$$



little more to identification

- Inequality constraints may or may not help identification since it limits the range of possible values of a free parameter, not to a particular value
- Constant values, equality, and linear combinations of other parameters are also a form of restriction that affects identification, but the identification rules considered so far assume only zero-constraints (except for t-rule)
- Identification discussed so far focuses only on whether the equation systems are solvable, which does not guarantee convergence or properness of a solution (e.g., negative ψ_{ii})
- Identification rules considered here applies to a part of the general model, not the “measurement models” part