Research Organization of Information and Systems
The Institute of Statistical Mathematics

An Introduction to Structural Equation Modeling With the sem Package in R

John Fox

McMaster University
Canada

November 2012
Tokyo, Japan

1. Introduction

- *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.
  - Indeed, two variables in a SEM can even effect one-another reciprocally, either directly, or indirectly through a "feedback" loop.
- Structural-equation models can include variables that are not measured directly, but rather indirectly through their effects (called *indicators*) or, sometimes, through their observable causes.
  - Unmeasured variables are variously termed *latent variables*, *constructs*, or *factors*.
- Modern structural-equation methods represent a confluence of work in many disciplines, including biostatistics, econometrics, psychometrics, and social statistics. The general synthesis of these various traditions dates to the late 1960s and early 1970s.

- This introduction to SEMs takes up several topics:
  - The form and specification of observed-variables SEMs.
  - Instrumental variables estimation.
  - The "identification problem": Determining whether or not a SEM, once specified, can be estimated.
  - Estimation of observed-variable SEMs.
  - Structural-equation models with latent variables, measurement errors, and multiple indicators.
  - The "LISREL" model: A general structural-equation model with latent variables.
- I will estimate SEMs using the **sem** package in R.
  - The current version of the **sem** package is joint work with Zhenghua Nie and Jarrett Brynes.

2. Some References

- J. Fox, "Linear Structural-Equation Models," Chapter 4, *Linear Statistical Models and Related Methods* (Wiley, 1984).
- J. Fox, "Structural-Equation Modeling with the sem Package in R," *Structural Equation Modeling*, 2006, 13:465-486 (out of date).
- J. Fox, "Structural Equation Modeling in R with the sem Package: An Appendix to An R Companion to Applied Regression, Second Edition, by John Fox and Sanford Weisberg," September 2012.
- K. A. Bollen, *Structural Equations with Latent Variables* (Wiley, 1989).
- K. A. Bollen, "Latent Variables in Psychology and the Social Sciences," *Annual Review of Psychology*, 2002, 53: 605-634.

3. Specification of Structural-Equation Models

- Structural-equation models 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.
  - Claiming that a relationship is causal based on observational data is no less problematic in a SEM than it is in a single-equation regression model.
  - Such a claim is intrinsically problematic and requires support beyond the data at hand.

- Several classes of variables appears in SEMs:
  - *Endogenous variables* are the response variables of the model.
    - There is one *structural equation* (regression equation) for each endogenous variable.
    - An endogenous variable may, however, also appear as an explanatory variable in other structural equations.
    - For the kinds of models that I will consider, the endogenous variables are (as in the single-equation linear model) quantitative continuous variables.
  - *Exogenous variables* appear only as explanatory variables in the structural equations.
    - The values of exogenous variable are therefore determined outside of the model (hence the term).
    - Like the explanatory variables in a linear model, exogenous variables are assumed to be measured without error (but see the later discussion of latent-variable models).

    - Exogenous variables can be categorical (represented, as in a linear model, by dummy regressors or other sorts of contrasts).
  - *Structural errors* (or *disturbances*) represent the aggregated omitted causes of the endogenous variables, along with measurement error (and possibly intrinsic randomness) in the endogenous variables.
    - There is one error variable for each endogenous variable (and hence for each structural equation).
    - The errors are assumed to have zero expectations and to be independent of (or at least uncorrelated with) the exogenous variables.
    - The errors for different observations are assumed to be independent of one another, but (depending upon the form of the model) different errors for the same observation may be related.

    - Each error variable is assumed to have constant variance across observations, although different error variables generally will have different variances (and indeed different units of measurement — the square units of the corresponding endogenous variables). As in a linear model, the assumption of constant error variance can be relaxed, though I will not pursue this possibility.
    - As in linear models, I will sometimes assume that the errors are normally distributed.
- I will use the following notation for writing down SEMs:
  - Endogenous variables: $y_k, y_{k'}$
  - Exogenous variables: $x_j, x_{j'}$
  - Errors: $\varepsilon_k, \varepsilon_{k'}$

- *Structural coefficients* (i.e., regression coefficients) representing the direct (partial) effect
  - of an exogenous on an endogenous variable, x_j on y_k: γ_{kj} (gamma).
    - Note that the subscript of the response variable comes first.
  - of an endogenous variable on another endogenous variable, $y_{k'}$ on y_k: $\beta_{kk'}$ (beta).
- Covariances between
  - two exogenous variables, x_j and $x_{j'}$: $\sigma_{jj'}$
  - two error variables, ε_k and $\varepsilon_{k'}$: $\sigma_{kk'}$
- When I require them, other covariances are represented similarly.
- Variances will be written either as σ_j^2 or as σ_{jj} (i.e., the covariance of a variable with itself), as is convenient.

3.1 Path Diagrams

- An intuitively appealing way of representing a SEM is in the form of a causal graph, called a path diagram. An example, from Duncan, Haller, and Portes's (1968) study of peer influences on the aspirations of high-school students, appears in Figure 1.
- The following conventions are used in the path diagram:
  - A directed (single-headed) arrow represents a direct effect of one variable on another; each such arrow is labelled with a structural coefficient.
  - A bidirectional (two-headed) arrow represents a covariance, between exogenous variables or between errors, that is not given causal interpretation.
  - I give each variable in the model (x, y , and ε) a unique subscript; I find that this helps to keep track of variables and coefficients.

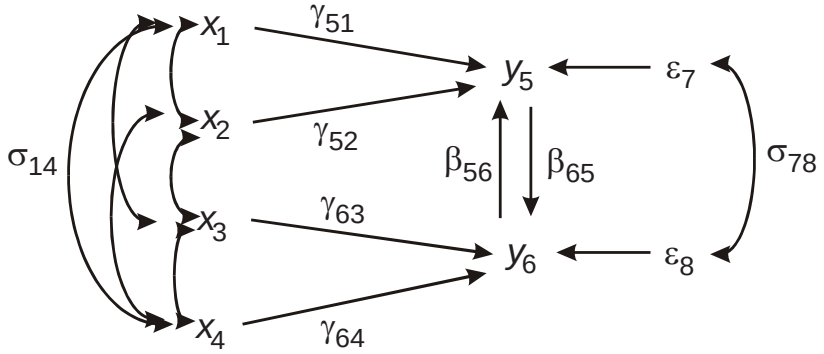

Figure 1. Duncan, Haller, and Portes's (nonrecursive) peer-influences model: x_1, respondent's IQ; x_2, respondent's family SES; x_3, best friend's family SES; x_4, best friend's IQ; y_5 , respondent's occupational aspiration; y_6, best friend's occupational aspiration. So as not to clutter the diagram, only one exogenous covariance, σ_{14}, is shown.

- When two variables are not linked by a directed arrow it *does not* necessarily mean that one does not affect the other:
  - For example, in the Duncan, Haller, and Portes model, respondent's IQ (x_1) can affect best friend's occupational aspiration (y_6), but only *indirectly*, through respondent's aspiration (y_5).
  - The absence of a directed arrow between respondent's IQ and best friend's aspiration means that there is no *partial relationship* between the two variables when the direct causes of best friend's aspiration are *held constant*.
  - In general, indirect effects can be identified with "compound paths" through the path diagram.

3.2 Structural Equations

- The structural equations of a model can be read straightforwardly from the path diagram.
  - For example, for the Duncan, Haller, and Portes peer-influences model:

$$y_{5i} = \gamma_{50} + \gamma_{51}x_{1i} + \gamma_{52}x_{2i} + \beta_{56}y_{6i} + \varepsilon_{7i}$$
$$y_{6i} = \gamma_{60} + \gamma_{63}x_{3i} + \gamma_{64}x_{4i} + \beta_{65}y_{5i} + \varepsilon_{8i}$$

  - I'll usually simplify the structural equations by
    (i) suppressing the subscript i for observation;
    (ii) expressing all xs and y s as deviations from their populations means (and, later, from their means in the sample).
  - Putting variables in mean-deviation form gets rid of the constant terms (here, γ_{50} and γ_{60}) from the structural equations (which are rarely of interest), and will simplify some algebra later on.

- Applying these simplifications to the peer-influences model:

$$y_5 = \gamma_{51}x_1 + \gamma_{52}x_2 + \beta_{56}y_6 + \varepsilon_7$$
$$y_6 = \gamma_{63}x_3 + \gamma_{64}x_4 + \beta_{65}y_5 + \varepsilon_8$$

3.3 Matrix Form of the Model

- It is sometimes helpful (e.g., for generality) to cast a structural-equation model in matrix form.
- To illustrate, I'll begin by rewriting the Duncan, Haller and Portes model, shifting all observed variables (i.e., with the exception of the errors) to the left-hand side of the model, and showing all variables explicitly; variables missing from an equation therefore get 0 coefficients, while the response variable in each equation is shown with a coefficient of 1:

$$1y_5 - \beta_{56}y_6 - \gamma_{51}x_1 - \gamma_{52}x_2 + 0x_3 + 0x_4 = \varepsilon_7$$
$$-\beta_{65}y_5 + 1y_6 + 0x_1 + 0x_2 - \gamma_{63}x_3 - \gamma_{64}x_4 = \varepsilon_8$$

- Collecting the endogenous variables, exogenous variables, errors, and coefficients into vectors and matrices, I can write

$$\begin{bmatrix} 1 & -\beta_{56} \\ -\beta_{65} & 1 \end{bmatrix} \begin{bmatrix} y_5 \\ y_6 \end{bmatrix} + \begin{bmatrix} -\gamma_{51} & -\gamma_{52} & 0 & 0 \\ 0 & 0 & -\gamma_{63} & -\gamma_{64} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \varepsilon_7 \\ \varepsilon_8 \end{bmatrix}$$

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

$$\underset{(q\times q)}{\mathbf{B}}\ \underset{(q\times 1)}{\mathbf{y}_i} + \underset{(q\times m)}{\mathbf{\Gamma}}\ \underset{(m\times 1)}{\mathbf{x}_i} = \underset{(q\times 1)}{\boldsymbol{\varepsilon}_i}$$

  - The $\mathbf{B}$ (Beta) and $\mathbf{\Gamma}$ (Gamma) matrices of structural coefficients typically contain some 0 elements, and the diagonal entries of the $\mathbf{B}$ matrix are 1s.

- I can also write the model for all n observations in the sample:

$$\underset{(n\times q)}{\mathbf{Y}}\ \underset{(q\times q)}{\mathbf{B}'} + \underset{(n\times m)}{\mathbf{X}}\ \underset{(m\times q)}{\mathbf{\Gamma}'} = \underset{(n\times q)}{\mathbf{E}}$$

  - I have transposed the structural-coefficient matrices $\mathbf{B}$ and $\mathbf{\Gamma}$, writing each structural equation as a column (rather than as a row), so that each observation comprises a row of the matrices $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{E}$ of endogenous variables, exogenous variables, and errors.

3.4 Recursive, Block-Recursive, and Nonrecursive Structural-Equation Models

- An important type of SEM, called a *recursive* model, has two defining characteristics:
  (a) Different error variables are independent (or, at least, uncorrelated).
  (b) Causation in the model is unidirectional: There are no reciprocal paths or feedback loops, as shown in Figure 2.
- Put another way, the $\mathbf{B}$ matrix for a recursive SEM is lower-triangular, while the error-covariance matrix $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ is diagonal.
- An illustrative recursive model, from Blau and Duncan's seminal monograph, *The American Occupational Structure* (1967), appears in Figure 3.
  - For the Blau and Duncan model:

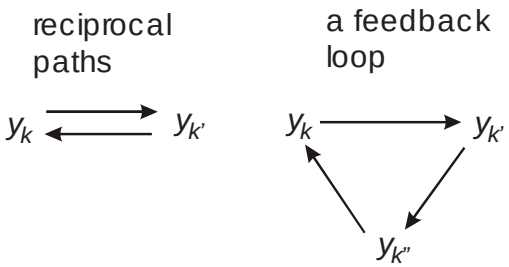

Figure 2. Reciprocal paths and feedback loops cannot appear in a recursive model.

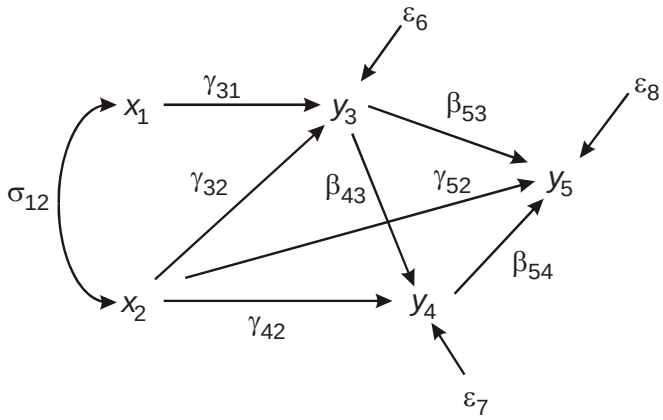

Figure 3. Blau and Duncan's "basic stratification" model: x_1, father's education; x_2, father's occupational status; y_3, respondent's (son's) education; y_4, respondent's first-job status; y_5, respondent's present (1962) occupational status.

$$\boldsymbol{\Gamma} = \begin{bmatrix} -\gamma_{31} & -\gamma_{32} \\ 0 & -\gamma_{42} \\ 0 & -\gamma_{52} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ -\beta_{43} & 1 & 0 \\ -\beta_{53} & -\beta_{54} & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \begin{bmatrix} \sigma_6^2 & 0 & 0 \\ 0 & \sigma_7^2 & 0 \\ 0 & 0 & \sigma_8^2 \end{bmatrix}$$

- Sometimes the requirements for unidirectional causation and independent errors are met by subsets ("blocks") of endogenous variables and their associated errors rather than by the individual variables. Such a model is called *block recursive*.
- An illustrative block-recursive model for the Duncan, Haller, and Portes peer-influences data is shown in Figure 4.

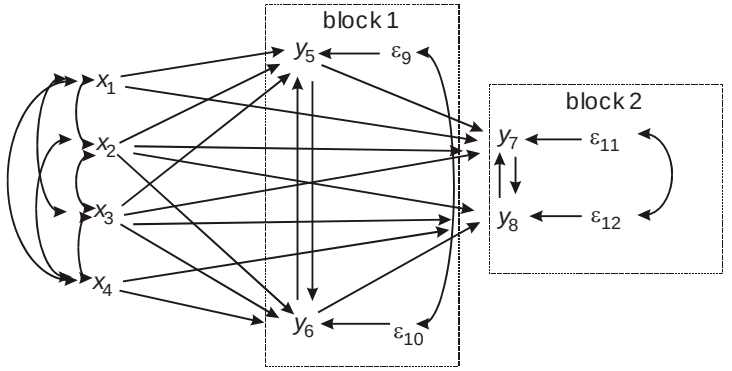

Figure 4. An extended, block-recursive model for Duncan, Haller, and Portes's peer-influences data: x_1, respondent's IQ; x_2, respondent's family SES; x_3, best friend's family SES; x_4, best friend's IQ; y_5, respondent's occupational aspiration; y_6, best friend's occupational aspiration; y_7, respondent's educational aspiration; y_8, best friend's educational aspiration.

- Here

$$\mathbf{B} = \left[\begin{array}{cc|cc} 1 & -\beta_{56} & 0 & 0 \\ -\beta_{65} & 1 & 0 & 0 \\ \hline -\beta_{75} & 0 & 1 & -\beta_{78} \\ 0 & -\beta_{86} & -\beta_{87} & 1 \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \left[\begin{array}{cc|cc} \sigma_9^2 & \sigma_{9,10} & 0 & 0 \\ \sigma_{10,9} & \sigma_{10}^2 & 0 & 0 \\ \hline 0 & 0 & \sigma_{11}^2 & \sigma_{11,12} \\ 0 & 0 & \sigma_{12,11} & \sigma_{12}^2 \end{array}\right] = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

- A model that is neither recursive nor block-recursive (such as the model for Duncan, Haller and Portes's data in Figure 1) is termed *nonrecursive*.

4. Instrumental-Variables Estimation*

- *Instrumental-variables (IV) estimation* is a method of deriving estimators that is useful for understanding whether estimation of a structural equation model is possible (the "identification problem") and for obtaining estimates of structural parameters when it is.

4.1 Simple Regression

- To understand the IV approach to estimation, consider first the following route to the *ordinary-least-squares* (*OLS*) estimator of the simple-regression model,

$$y = \beta x + \varepsilon$$

  where the variables x and y are in mean-deviation form, eliminating the regression constant from the model; that is, $E(y) = E(x) = 0$.
  - By the usual assumptions of this model, $E(\varepsilon) = 0$; Var$(\varepsilon) = \sigma_\varepsilon^2$; and x, ε are independent.

* If necessasry.

- Now multiply both sides of the model by x and take expectations:

$$\begin{aligned} xy &= \beta x^2 + x\varepsilon \\ E(xy) &= \beta E(x^2) + E(x\varepsilon) \\ \mathsf{Cov}(x, y) &= \beta \mathsf{Var}(x) + \mathsf{Cov}(x\varepsilon) \\ \sigma_{xy} &= \beta\sigma_x^2 + 0 \end{aligned}$$

  where $\mathsf{Cov}(x\varepsilon) = 0$ because x and ε are independent.

- Solving for the regression coefficient β,

$$\beta = \frac{\sigma_{xy}}{\sigma_x^2}$$

- Of course, we don't know the population covariance of x and y, nor do we know the population variance of x, but we can estimate both of these parameters consistently:

$$\begin{aligned} s_x^2 &= \frac{\sum(x_i - \overline{x})^2}{n-1} \\ s_{xy} &= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1} \end{aligned}$$

  In these formulas, the variables are expressed in raw-score form, and so I show the subtraction of the sample means explicitly.

- A consistent estimator of β is then

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

  which we recognize as the OLS estimator.

▶ Imagine, alternatively, that x and ε are not independent, but that ε is independent of some other variable z.

- Suppose further that z and x are correlated — that is, $\mathsf{Cov}(x, z) \neq 0$.
- Then, proceeding as before, but multiplying through by z rather than by x (with all variable expressed as deviations from their expectations):

$$\begin{aligned} zy &= \beta zx + z\varepsilon \\ E(zy) &= \beta E(zx) + E(z\varepsilon) \\ \mathsf{Cov}(z, y) &= \beta \mathsf{Cov}(z, x) + \mathsf{Cov}(z\varepsilon) \\ \sigma_{zy} &= \beta\sigma_{zx} + 0 \\ \beta &= \frac{\sigma_{zy}}{\sigma_{zx}} \end{aligned}$$

  where $\mathsf{Cov}(z\varepsilon) = 0$ because z and ε are independent.

- Substituting sample for population covariances gives the *instrumental variables estimator* of β:

$$b_{\mathsf{IV}} = \frac{s_{zy}}{s_{zx}} = \frac{\sum(z_i - \overline{z})(y_i - \overline{y})}{\sum(z_i - \overline{z})(x_i - \overline{x})}$$

– The variable z is called an *instrumental variable* (or, simply, an *instrument*).

– b_{IV} is a consistent estimator of the population slope β, because the sample covariances s_{zy} and s_{zx} are consistent estimators of the corresponding population covariances σ_{zy} and σ_{zx}.

4.2 Multiple Regression

- The generalization to multiple-regression models is straightforward.
  - For example, for a model with two explanatory variables,

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

(with x_1, x_2, and y all expressed as deviations from their expectations).

  - If we can assume that the error ε is independent of x_1 and x_2, then we can derive the population analog of estimating equations by multiplying through by the two explanatory variables in turn, obtaining

$$\begin{aligned} E(x_1 y) &= \beta_1 E(x_1^2) + \beta_2 E(x_1 x_2) + E(x_1 \varepsilon) \\ E(x_2 y) &= \beta_1 E(x_1 x_2) + \beta_2 E(x_2^2) + E(x_2 \varepsilon) \end{aligned}$$

$$\begin{aligned} \sigma_{x_1 y} &= \beta_1 \sigma^2_{x_1} + \beta_2 \sigma_{x_1 x_2} + 0 \\ \sigma_{x_2 y} &= \beta_1 \sigma_{x_1 x_2} + \beta_2 \sigma^2_{x_2} + 0 \end{aligned}$$

  - Substituting sample for population variances and covariances produces the OLS estimating equations:

$$\begin{aligned} s_{x_1 y} &= b_1 s^2_{x_1} + b_2 s_{x_1 x_2} \\ s_{x_2 y} &= b_1 s_{x_1 x_2} + b_2 s^2_{x_2} \end{aligned}$$

  - Alternatively, if we cannot assume that ε is independent of the xs, but can assume that ε is independent of two other variables, z_1 and z_2, then

$$\begin{aligned} E(z_1 y) &= \beta_1 E(z_1 x_1) + \beta_2 E(z_1 x_2) + E(z_1 \varepsilon) \\ E(z_2 y) &= \beta_1 E(z_2 x_1) + \beta_2 E(z_2 x_2) + E(z_2 \varepsilon) \end{aligned}$$

$$\begin{aligned} \sigma_{z_1 y} &= \beta_1 \sigma_{z_1 x_1} + \beta_2 \sigma_{z_1 x_2} + 0 \\ \sigma_{z_2 y} &= \beta_1 \sigma_{z_2 x_1} + \beta_2 \sigma_{z_2 x_2} + 0 \end{aligned}$$

  - the IV estimating equations are obtained by the now familiar step of substituting consistent sample estimators for the population covariances:

$$\begin{aligned} s_{z_1 y} &= b_1 s_{z_1 x_1} + b_2 s_{z_1 x_2} \\ s_{z_2 y} &= b_1 s_{z_2 x_1} + b_2 s_{z_2 x_2} \end{aligned}$$

  - For the IV estimating equations to have a unique solution, it's necessary that there not be an analog of perfect collinearity.
    – For example, neither x_1 nor x_2 can be uncorrelated with both z_1 and z_2.

- Good instrumental variables, while remaining uncorrelated with the error, should be as correlated as possible with the explanatory variables.
  - In this context, 'good' means yielding relatively small coefficient standard errors (i.e., producing efficient estimates).

- OLS is a special case of IV estimation, where the instruments and the explanatory variables are one and the same.
  – When the explanatory variables are uncorrelated with the error, the explanatory variables are their own best instruments, since they are perfectly correlated with themselves.
  – Indeed, the Gauss-Markov theorem insures that when it is applicable, the OLS estimator is the best (i.e., minimum variance or most efficient) linear unbiased estimator (*BLUE*).

4.3 Instrumental-Variables Estimation in Matrix Form

- Our object is to estimate the model

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}}\underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2_\varepsilon \mathbf{I}_n)$.

- Of course, if $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are independent, then we can use the OLS estimator

$$\mathbf{b}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

with estimated covariance matrix

$$\widehat{V}(\mathbf{b}_{\text{OLS}}) = s^2_{\text{OLS}}(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s^2_{\text{OLS}} = \frac{\mathbf{e}'_{\text{OLS}}\mathbf{e}_{\text{OLS}}}{n-k-1}$$

for

$$\mathbf{e}_{\text{OLS}} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{OLS}}$$

- Suppose, however, that we cannot assume that $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are independent, but that we have observations on $k+1$ instrumental variables, $\underset{(n\times k+1)}{\mathbf{Z}}$, that are independent of $\boldsymbol{\varepsilon}$.
  - For greater generality, I have not put the variables in mean-deviation form, and so the model includes a constant; the matrices $\mathbf{X}$ and $\mathbf{Z}$ therefore each include an initial column of ones.
  - A development that parallels the previous scalar treatment leads to the IV estimator

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

with estimated covariance matrix

$$\widehat{V}(\mathbf{b}_{\text{IV}}) = s^2_{\text{IV}}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}$$

where

$$s^2_{\text{IV}} = \frac{\mathbf{e}'_{\text{IV}}\mathbf{e}_{\text{IV}}}{n-k-1}$$

for

$$\mathbf{e}_{\text{IV}} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{IV}}$$

- Since the results for IV estimation are asymptotic, I could also estimate the error variance with n rather than $n-k-1$ in the denominator, but dividing by degrees of freedom produces a larger variance estimate and hence is conservative.
- For $\mathbf{b}_{\text{IV}}$ to be unique $\mathbf{Z}'\mathbf{X}$ must be nonsingular (just as $\mathbf{X}'\mathbf{X}$ must be nonsingular for the OLS estimator).

5. The Identification Problem

- If a parameter in a structural-equation model can be estimated then the parameter is said to be *identified*; otherwise, it is *underidentified* (or *unidentified*).
  - If all of the parameters in a structural equation are identified, then so is the equation.
  - If all of the equations in a SEM are identified, then so is the model.
  - Structural equations and models that are not identified are also termed underidentified.
- If only one estimate of a parameter is available, then the parameter is *just-identified* or *exactly identified*.
- If more than one estimate is available, then the parameter is *overidentified*.

- The same terminology extends to structural equations and to models: An identified structural equation or SEM with one or more overidentified parameters is itself overidentified.
- Establishing whether a SEM is identified is called the *identification problem*.
  - Identification is usually established one structural equation at a time.

5.1 Identification of Nonrecursive Models: The Order Condition

- Using instrumental variables, I can derive a necessary (but, as it turns out, not sufficient) condition for identification of nonrecursive models called the *order condition*.
  - Because the order condition is not sufficient to establish identification, it is possible (though rarely the case) that a model can meet the order condition but not be identified.
  - There is a necessary and sufficient condition for identification called the *rank condition*, which I will not develop here. The rank condition is described in the references.

- The terms “order condition” and “rank condition” derive from the order (number of rows and columns) and rank (number of linearly independent rows and columns) of a matrix that can be formulated during the process of identifying a structural equation. I will not pursue this approach.
- Both the order and rank conditions apply to nonrecursive models without restrictions on disturbance covariances.
  - Such restrictions can sometimes serve to identify a model that would not otherwise be identified.
  - More general approaches are required to establish the identification of models with disturbance-covariance restrictions. Again, these are taken up in the references.
  - I will, however, use the IV approach to consider the identification of two classes of models with restrictions on disturbance covariances: recursive and block-recursive models.

- The order condition is best developed from an example.
  - Recall the Duncan, Haller, and Portes peer-influences model, reproduced in Figure 5.
  - Let us focus on the first of the two structural equations of the model,
$$y_5 = \gamma_{51}x_1 + \gamma_{52}x_2 + \beta_{56}y_6 + \varepsilon_7$$
where all variables are expressed as deviations from their expectations.
    - There are three structural parameters to estimate in this equation, γ_{51}, γ_{52}, and β_{56}.
  - It would be inappropriate to perform OLS regression of y_5 on x_1, x_2, and y_6 to estimate this equation, because we cannot reasonably assume that the endogenous explanatory variable y_6 is uncorrelated with the error ε_7.
    - ε_7 may be correlated with ε_8, which is one of the components of y_6 .
    - ε_7 is a component of y_5 which is a cause (as well as an effect) of y_6.

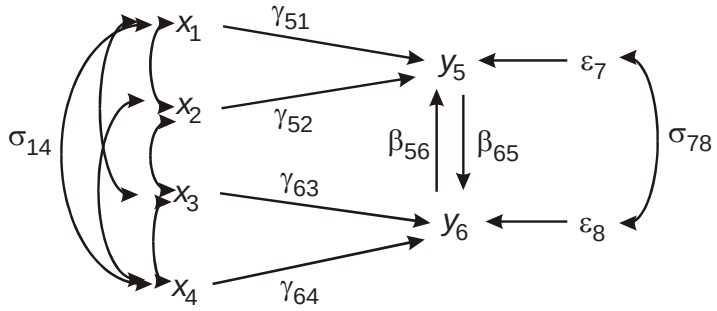

Figure 5. Duncan, Haller, and Portes nonrecursive peer-influences model (repeated).

- This conclusion is more general: we cannot assume that endogenous explanatory variables are uncorrelated with the error of a structural equation.
  - As we will see, however, we *will* be able to make this assumption in recursive models.
- Nevertheless, we can use the four exogenous variables x_1, x_2, x_3, and x_4, as instrumental variables to obtaining estimating equations for the structural equation:
  - For example, multiplying through the structural equation by x_1 and taking expectations produces

$$\begin{aligned} x_1 y_5 &= \gamma_{51} x_1^2 + \gamma_{52} x_1 x_2 + \beta_{56} x_1 y_6 + x_1 \varepsilon_7 \\ E(x_1 y_5) &= \gamma_{51} E(x_1^2) + \gamma_{52} E(x_1 x_2) + \beta_{56} E(x_1 y_6) + E(x_1 \varepsilon_7) \\ \sigma_{15} &= \gamma_{51} \sigma_1^2 + \gamma_{52} \sigma_{12} + \beta_{56} \sigma_{16} + 0 \end{aligned}$$

  since $\sigma_{17} = E(x_1 \varepsilon_7) = 0$.

– Applying all four exogenous variables,

IV	Estimating Equation
x_1	$\sigma_{15} = \gamma_{51}\sigma_1^2 + \gamma_{52}\sigma_{12} + \beta_{56}\sigma_{16}$
x_2	$\sigma_{25} = \gamma_{51}\sigma_{12} + \gamma_{52}\sigma_2^2 + \beta_{56}\sigma_{26}$
x_3	$\sigma_{35} = \gamma_{51}\sigma_{13} + \gamma_{52}\sigma_{23} + \beta_{56}\sigma_{36}$
x_4	$\sigma_{45} = \gamma_{51}\sigma_{14} + \gamma_{52}\sigma_{24} + \beta_{56}\sigma_{46}$

- If the model is correct, then all of these equations, involving population variances, covariances, and structural parameters, hold simultaneously and exactly.
- If we had access to the population variances and covariances, then, we could solve for the structural coefficients γ_{51}, γ_{52}, and β_{56} even though there are four equations and only three parameters.
- Since the four equations hold simultaneously, we could obtain the solution by eliminating any one and solving the remaining three.

- Translating from population to sample produces four IV estimating equations for the three structural parameters:

$$\begin{aligned} s_{15} &= \widehat{\gamma}_{51} s_1^2 + \widehat{\gamma}_{52} s_{12} + \widehat{\beta}_{56} s_{16} \\ s_{25} &= \widehat{\gamma}_{51} s_{12} + \widehat{\gamma}_{52} s_2^2 + \widehat{\beta}_{56} s_{26} \\ s_{35} &= \widehat{\gamma}_{51} s_{13} + \widehat{\gamma}_{52} s_{23} + \widehat{\beta}_{56} s_{36} \\ s_{45} &= \widehat{\gamma}_{51} s_{14} + \widehat{\gamma}_{52} s_{24} + \widehat{\beta}_{56} s_{46} \end{aligned}$$

  - The s_j^2s and $s_{jj'}$s are sample variances and covariances that can be calculated directly from sample data, while $\widehat{\gamma}_{51}$, $\widehat{\gamma}_{52}$, and $\widehat{\beta}_{56}$ are estimates of the structural parameters, for which we want to solve the estimating equations.
  - There is a problem, however: The four estimating equations in the three unknown parameter estimates will not hold precisely:
    - Because of sampling variation, there will be no set of estimates that simultaneously satisfies the four estimating equations.

  - That is, the four estimating equations in three unknown parameters are *overdetermined*.
  - Under these circumstances, the three parameters and the structural equation are said to be *overidentified*.
- It is important to appreciate the nature of the problem here:
  - We have *too much* rather than too little information.
  - We could simply throw away one of the four estimating equations and solve the remaining three for consistent estimates of the structural parameters.
  - The estimates that we would obtain would depend, however, on which estimating equation was discarded.
  - Moreover, throwing away an estimating equation, while yielding consistent estimates, discards information that could be used to improve the *efficiency* of estimation.

- To illuminate the nature of overidentification, consider the following, even simpler, example:
  - We want to estimate the structural equation
$$y_5 = \gamma_{51}x_1 + \beta_{54}y_4 + \varepsilon_6$$
and have available as instruments the exogenous variables x_1, x_2, and x_3.
  - Then, in the population, the following three equations hold simultaneously:

IV	Estimating Equation
x_1	$\sigma_{15} = \gamma_{51}\sigma_1^2 + \beta_{54}\sigma_{14}$
x_2	$\sigma_{25} = \gamma_{51}\sigma_{12} + \beta_{54}\sigma_{24}$
x_3	$\sigma_{35} = \gamma_{51}\sigma_{13} + \beta_{54}\sigma_{34}$

  - These linear equations in the parameters γ_{51} and β_{54} are illustrated in Figure 6 (a), which is constructed assuming particular values for the population variances and covariances in the equations.

- The important aspect of this illustration is that the three equations intersect at a single point, determining the structural parameters, which are the solution to the equations.
- The three estimating equations are
$$\begin{aligned} s_{15} &= \widehat{\gamma}_{51}s_1^2 + \widehat{\beta}_{54}s_{14} \\ s_{25} &= \widehat{\gamma}_{51}s_{12} + \widehat{\beta}_{54}s_{24} \\ s_{35} &= \widehat{\gamma}_{51}s_{13} + \widehat{\beta}_{54}s_{34} \end{aligned}$$
- As illustrated in Figure 6 (b), because the sample variances and covariances are not exactly equal to the corresponding population values, the estimating equations do not in general intersect at a common point, and therefore have no solution.
- Discarding an estimating equation, however, produces a solution, since each pair of lines intersects at a point.

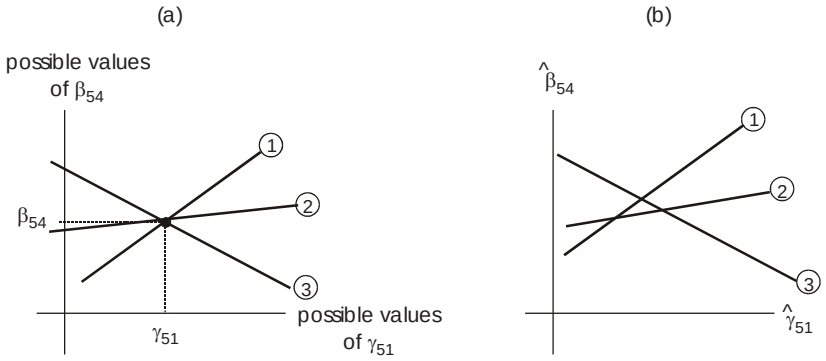

Figure 6. Population equations (a) and corresponding estimating equations (b) for an overidentified structural equation with two parameters and three estimating equations. The population equations have a solution for the parameters, but the estimating equations do not.

- Let us return to the Duncan, Haller, and Portes model, and add a path from x_3 to y_5, so that the first structural equation becomes

$$y_5 = \gamma_{51}x_1 + \gamma_{52}x_2 + \gamma_{53}x_3 + \beta_{56}y_6 + \varepsilon_7$$

  - There are now four parameters to estimate (γ_{51}, γ_{52}, γ_{53}, and β_{56}), and four IVs (x_1, x_2, x_3, and x_4), which produces four estimating equations.
  - With as many estimating equations as unknown structural parameters, there is only one way of estimating the parameters, which are therefore *just identified*.
  - We can think of this situation as a kind of balance sheet with IVs as "credits" and structural parameters as "debits."

  - For a just-identified structural equation, the numbers of credits and debits are the same:

Credits IVs	*Debits* parameters
x_1	γ_{51}
x_2	γ_{52}
x_3	γ_{53}
x_4	β_{56}
4	4

- In the original specification of the Duncan, Haller, and Portes model, there were only three parameters in the first structural equation, producing a surplus of IVs, and an overidentified structural equation:

Credits IVs	*Debits* parameters
x_1	γ_{51}
x_2	γ_{52}
x_3	β_{56}
x_4	
4	3

- Now let us add still another path to the model, from x_4 to y_5, so that the first structural equation becomes

$$y_5 = \gamma_{51}x_1 + \gamma_{52}x_2 + \gamma_{53}x_3 + \gamma_{54}x_4 + \beta_{56}y_6 + \varepsilon_7$$

  - Now there are fewer IVs available than parameters to estimate in the structural equation, and so the equation is *underidentified*:

Credits IVs	*Debits* parameters
x_1	γ_{51}
x_2	γ_{52}
x_3	γ_{53}
x_4	γ_{54}
	β_{56}
4	5

  - That is, we have only four estimating equations for five unknown parameters, producing an underdetermined system of estimating equations.

- From these examples, we can abstract the *order condition for identification* of a structural equation: For the structural equation to be identified, we need at least as many exogenous variables (instrumental variables) as there are parameters to estimate in the equation.
  - Since structural equation models have more than one endogenous variable, the order condition implies that some potential explanatory variables must be excluded apriori from each structural equation of the model for the model to be identified.
  - Put another way, for each endogenous explanatory variable in a structural equation, at least one exogenous variable must be excluded from the equation.
  - Suppose that there are m exogenous variable in the model:
    - A structural equation with fewer than m structural parameters is overidentified.
    - A structural equation with exactly m structural parameters is just-identified.

  - A structural equation with more than m structural parameters is underidentified, and cannot be estimated.

5.2 Identification of Recursive and Block-Recursive Models†

- The pool of IVs for estimating a structural equation in a recursive model includes not only the exogenous variables but *prior endogenous variables* as well.
  - Because the explanatory variables in a structural equation are drawn from among the exogenous and prior endogenous variables, there will always be at least as many IVs as there are explanatory variables (i.e., structural parameters to estimate).
  - Consequently, structural equations in a recursive model are necessarily identified.
- To understand this result, consider the Blau and Duncan basic-stratification model, reproduced in Figure 7.

† As time permits.

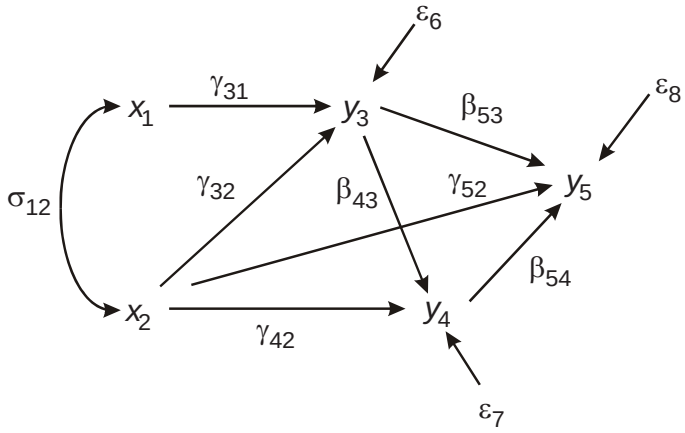

Figure 7. Blau and Duncan's recursive basic-stratification model (repeated).

- The first structural equation of the model is

$$y_3 = \gamma_{31}x_1 + \gamma_{32}x_2 + \varepsilon_6$$

with "balance sheet"

Credits IVs	*Debits* parameters
x_1	γ_{31}
x_2	γ_{32}
2	2

  - Because there are equal numbers of IVs and structural parameters, the first structural equation is just-identified.

  - More generally, the first structural equation in a recursive model can have only exogenous explanatory variables (or it wouldn't be the *first* equation).
    - If all the exogenous variables appear as explanatory variables (as in the Blau and Duncan model), then the first structural equation is just-identified.
    - If any exogenous variables are excluded as explanatory variables from the first structural equation, then the equation is overidentified.
- The second structural equation in the Blau and Duncan model is

$$y_4 = \gamma_{42}x_2 + \beta_{43}y_3 + \varepsilon_7$$

  - As before, the exogenous variable x_1 and x_2 can serve as IVs.
  - The prior endogenous variable y_3 can also serve as an IV, because (according to the first structural equation), y_3 is a linear combination of variables (x_1, x_2, and ε_6) that are all uncorrelated with the error ε_7 ($x_1$and x_2 because they are exogenous, ε_6 because it is another error variable).

  - The balance sheet is therefore

Credits IVs	*Debits* parameters
x_1	γ_{42}
x_2	β_{43}
y_3	
3	2

  - Because there is a surplus of IVs, the second structural equation is overidentified.
  - More generally, the second structural equation in a recursive model can have only the exogenous variables and the first (i.e., prior) endogenous variable as explanatory variables.

  - All of these *predetermined variables* are also eligible to serve as IVs.
  - If all of the predetermined variables appear as explanatory variables, then the second structural equation is just-identified; if any are excluded, the equation is overidentified.
- The situation with respect to the third structural equation is similar:

$$y_5 = \gamma_{52}x_2 + \beta_{53}y_3 + \beta_{54}y_4 + \varepsilon_8$$

  - Here, the eligible instrumental variables include (as always) the exogenous variables (x_1, x_2) and the two prior endogenous variables:
    - y_3 because it is a linear combination of exogenous variables (x_1 and x_2) and an error variable (ε_6), all of which are uncorrelated with the error from the third equation, e_8.
    - y_4 because it is a linear combination of variables (x_2, y_3, and ε_7 — as specified in the second structural equation), which are also all uncorrelated with ε_8.

– The balance sheet for the third structural equation indicates that the equation is overidentified:

Credits IVs	*Debits* parameters
x_1	γ_{52}
x_2	β_{53}
y_3	β_{54}
y_4	
4	3

- More generally:
  - All prior variables, including exogenous and prior endogenous variables, are eligible as IVs for estimating a structural equation in a recursive model.
  - If all of these prior variables also appear as explanatory variables in the structural equation, then the equation is just-identified.
  - If, alternatively, one or more prior variables are excluded, then the equation is overidentified.
  - A structural equation in a recursive model *cannot* be underidentified.

▶ A slight complication: There may only be a partial ordering of the endogenous variables.
- Consider, for example, the model in Figure 8.
  - This is a version of Blau and Duncan's model in which the path from y_3 to y_4 has been removed.
  - As a consequence, y_3 is no longer prior to y_4 in the model — indeed, the two variables are unordered.
  - Because the errors associated with these endogenous variables, ε_6 and ε_7, are uncorrelated with each other, however, y_3 is *still* available for use as an IV in estimating the equation for y_4.
  - Moreover, now y_4 is also available for use as an IV in estimating the equation for y_3, so the situation with respect to identification has, if anything, improved.

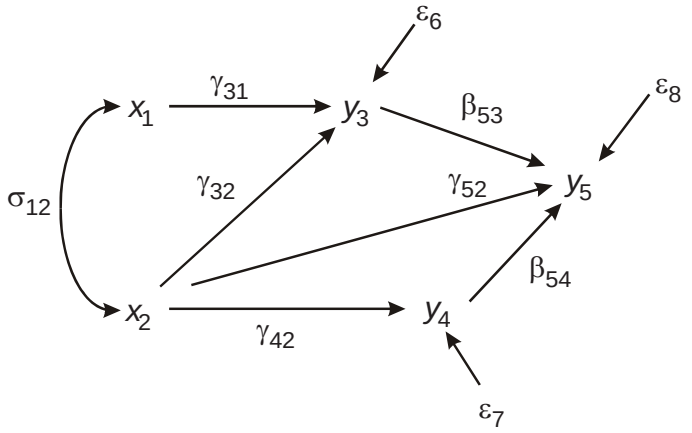

Figure 8. A recursive model (a modification of Blau and Duncan's model) in which there are two endogenous variables, y_3 and y_4, that are not ordered.

- In a block-recursive model, all exogenous variables and endogenous variables in prior blocks are available for use as IVs in estimating the structural equations in a particular block.
  - A structural equation in a block-recursive model may therefore be under-, just-, or overidentified, depending upon whether there are fewer, the same number as, or more IVs than parameters.
  - For example, recall the block-recursive model for Duncan, Haller, and Portes's peer-influences data, reproduced in Figure 9.
    - There are four IVs available to estimate the structural equations in the first block (for endogenous variables y_5 and y_6) — the exogenous variables (x_1, x_2, x_3, and x_4).
      - Because each of these structural equations has four parameters to estimate, each equation is just-identified.

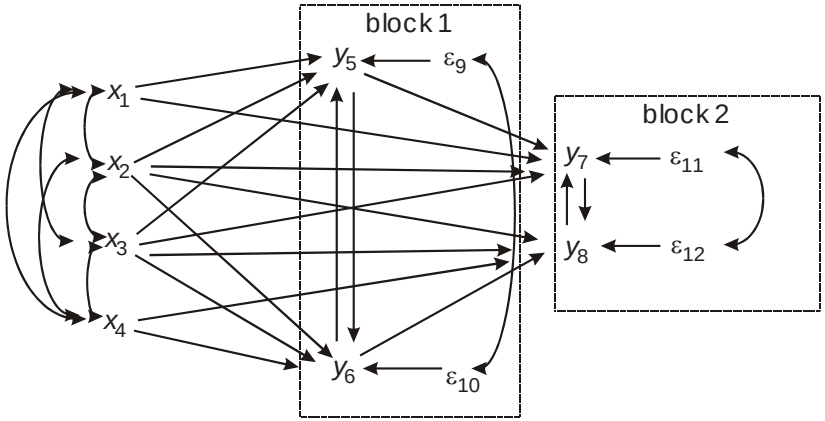

Figure 9. Block-recursive model for Duncan, Hallter and Portes's peer-influences data (repeated).

- There are six IVs available to estimate the structural equations in the second block (for endogenous variables y_7 and y_8) — the four exogenous variables plus the two endogenous variables (y_5 and y_6) from the first block.
  - Because each structural equation in the second block has five structural parameters to estimate, each equation is overidentified.
  - In the absence of the block-recursive restrictions on the disturbance covariances, only the exogenous variables would be available as IVs to estimate the structural equations in the second block, and these equations would consequently be underidentified.

6. Estimation of Structural-Equation Models

6.1 Estimating Nonrecursive Models

- There are two general and many specific approaches to estimating SEMs:
  (a) *Single-equation* or *limited-information* methods estimate each structural equation individually.
    - I will describe a single-equation method called *two-stage least squares* (*2SLS*).
    - Unlike OLS, which is also a limited-information method, 2SLS produces consistent estimates in nonrecursive SEMs.
    - Unlike direct IV estimation, 2SLS handles overidentified structural equations in a non-arbitrary manner.

– 2SLS also has a reasonable intuitive basis and appears to perform well — it is generally considered the best of the limited-information methods.

(b) *Systems* or *full-information* methods estimate all of the parameters in the structural-equation model simultaneously, including error variances and covariances.

– I will briefly describe a method called *full-information maximum-likelihood* (*FIML*).

– Full information methods are asymptotically more efficient than single-equation methods, although in a model with a misspecified equation, they tend to proliferate the specification error throughout the model.

– FIML appears to be the best of the full-information methods.

- Both 2SLS and FIML are implemented in the **sem** package for R.
  - *A note on terminology:* In the newer SEM literature, the term "FIML" is often reserved for full-information maximum-likelihood estimation in the presence of missing data, and the **sem** packages adopts this terminology. What I'm calling "FIML" for nonrecursive models in these slides is called "ML" in the package.

6.1.1 Two-Stage Least Squares

- Underidentified structural equations cannot be estimated.
- Just-identified equations can be estimated by direct application of the available IVs.
  - We have as many estimating equations as unknown parameters.
- For an overidentified structural equation, we have more than enough IVs.
  - There is a surplus of estimating equations which, in general, are not satisfied by a common solution.
  - 2SLS is a method for reducing the IVs to just the right number — but by combining IVs rather than discarding some altogether.

- Recall the first structural equation from Duncan, Haller, and Portes's peer-influences model:

$$y_5 = \gamma_{51}x_1 + \gamma_{52}x_2 + \beta_{56}y_6 + \varepsilon_7$$

  - This equation is overidentified because there are four IVs available (x_1, x_2, x_3, and x_4) but only three structural parameters to estimate (γ_{51}, γ_{52}, and β_{56}).
  - An IV must be correlated with the explanatory variables but uncorrelated with the error.
  - A good IV must be as correlated as possible with the explanatory variables, to produce estimated structural coefficients with small standard errors.
  - 2SLS chooses IVs by examining each explanatory variable in turn:
    – The *exogenous* explanatory variables x_1 and x_2 are their own best instruments because each is perfectly correlated with itself.

– To get a best IV for the *endogenous* explanatory variable y_6, we first regress this variable on all of the exogenous variables (by OLS), according to the *reduced-form model*

$$y_6 = \pi_{61}x_1 + \pi_{62}x_2 + \pi_{63}x_3 + \pi_{64}x_4 + \delta_6$$

producing fitted values

$$\widehat{y}_6 = \widehat{\pi}_{61}x_1 + \widehat{\pi}_{62}x_2 + \widehat{\pi}_{63}x_3 + \widehat{\pi}_{64}x_4$$

– Because $\widehat{y}_6$ is a linear combination of the xs — indeed, the linear combination most highly correlated with y_6 — it is (asymptotically) uncorrelated with the structural error ε_7.

– This is the *first stage* of 2SLS.

- Now we have just the right number of IVs: x_1, x_2, and $\widehat{y}_6$, producing three estimating equations for the three unknown structural parameters:

IV	2SLS Estimating Equation
x_1	$s_{15} = \widehat{\gamma}_{51}s_1^2 + \widehat{\gamma}_{52}s_{12} + \widehat{\beta}_{56}s_{16}$
x_2	$s_{25} = \widehat{\gamma}_{51}s_{12} + \widehat{\gamma}_{52}s_2^2 + \widehat{\beta}_{56}s_{26}$
$\widehat{y}_6$	$s_{5\widehat{6}} = \widehat{\gamma}_{51}s_{1\widehat{6}} + \widehat{\gamma}_{52}s_{2\widehat{6}} + \widehat{\beta}_{56}s_{6\widehat{6}}$

where, e.g., $s_{5\widehat{6}}$ is the sample covariance between y_5 and $\widehat{y}_6$.

- The generalization of 2SLS from this example is straightforward:
  - *Stage 1:* Regress each of the endogenous explanatory variables in a structural equation on all of the exogenous variables in the model, obtaining fitted values.
  - *Stage 2:* Use the fitted endogenous explanatory variables from stage 1 along with the exogenous explanatory variables as IVs to estimate the structural equation.
- If a structural equation is just-identified, then the 2SLS estimates are identical to those produced by direct application of the exogenous variables as IVs.

- There is an alternative route to the 2SLS estimator which, in the second stage, replaces each endogenous explanatory variable in the structural equation with the fitted values from the first stage regression, and then performs an OLS regression.
  - The second-stage OLS regression produces the same estimates as the IV approach.
  - The name “two-stage least squares” originates from this alternative approach.

- The 2SLS estimator for the jth structural equation in a nonrecursive model can be formulated in matrix form as follows:
  - Write the jth structural equation as

$$\underset{(n\times 1)}{\mathbf{y}_j} = \underset{(n\times q_j)(q_j\times 1)}{\mathbf{Y}_j\,\boldsymbol{\beta}_j} + \underset{(n\times m_j)(m_j\times 1)}{\mathbf{X}_j\,\boldsymbol{\gamma}_j} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}_j}$$

$$= [\mathbf{Y}_j, \mathbf{X}_j]\begin{bmatrix}\boldsymbol{\beta}_j\\ \boldsymbol{\gamma}_j\end{bmatrix} + \boldsymbol{\varepsilon}_j$$

where

$\mathbf{y}_j$ is the response-variable vector in structural equation j

$\mathbf{Y}_j$ is the matrix of q_j endogenous explanatory variables in equation j

$\boldsymbol{\beta}_j$ is the vector of structural parameters for the endogenous explanatory variables

$\mathbf{X}_j$ is the matrix of m_j exogenous explanatory variables in equation j, normally including a column of 1s

$\boldsymbol{\gamma}_j$ is the vector of structural parameters for the exogenous explanatory variables

$\boldsymbol{\varepsilon}_j$ is the error vector for structural equation j

- In the first stage of 2SLS, the endogenous explanatory variables are regressed on all m exogenous variables in the model, obtaining the OLS estimates of the reduced-form regression coefficients
$$\mathbf{P}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j$$
and fitted values
$$\widehat{\mathbf{Y}}_j = \mathbf{X}\mathbf{P}_j = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j$$
- In the second stage of 2SLS, we apply $\mathbf{X}_j$ and $\widehat{\mathbf{Y}}_j$ as instruments to the structural equation to obtain (after quite a bit of manipulation)
$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_j \\ \widehat{\boldsymbol{\gamma}}_j \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j & \mathbf{Y}_j'\mathbf{X}_j \\ \mathbf{X}_j'\mathbf{Y}_j & \mathbf{X}_j'\mathbf{X}_j \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_j \\ \mathbf{X}_j'\mathbf{y}_j \end{bmatrix}$$

- The estimated variance-covariance matrix of the 2SLS estimates is
$$\widehat{V}\begin{bmatrix} \widehat{\boldsymbol{\beta}}_j \\ \widehat{\boldsymbol{\gamma}}_j \end{bmatrix} = s^2_{e_j}\begin{bmatrix} \mathbf{Y}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j & \mathbf{Y}_j'\mathbf{X}_j \\ \mathbf{X}_j'\mathbf{Y}_j & \mathbf{X}_j'\mathbf{X}_j \end{bmatrix}^{-1}$$
where
$$s^2_{e_j} = \frac{\mathbf{e}_j'\mathbf{e}_j}{n - q_j - m_j}$$
$$\mathbf{e}_j = \mathbf{y}_j - \mathbf{Y}_j\widehat{\boldsymbol{\beta}}_j - \mathbf{X}_j\widehat{\boldsymbol{\gamma}}_j$$

6.1.2 Full-Information Maximum Likelihood

- Along with the 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
$$\log_e L(\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{\varepsilon\varepsilon}) = n\log_e|\det(\mathbf{B})| - \frac{nq}{2}\log_e 2\pi - \frac{n}{2}\log_e\det(\boldsymbol{\Sigma}_{\varepsilon\varepsilon})$$
$$-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{B}\mathbf{y}_i+\boldsymbol{\Gamma}\mathbf{x}_i)'\,\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}\,(\mathbf{B}\mathbf{y}_i+\boldsymbol{\Gamma}\mathbf{x}_i)$$
where $\det$ represents the determinant.

- The FIML estimates are the values of the parameters that maximize the likelihood under the constraints placed on the model – for example, that certain entries of $\mathbf{B}$, $\boldsymbol{\Gamma}$, and (possibly) $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ are 0.
- Estimated variances and covariances for the parameters are obtained from the inverse of the information matrix — the negative of the Hessian matrix of second-order partial derivatives of the log-likelihood — evaluated at the parameter estimates.
- The full general machinery of maximum-likelihood estimation is available — for example, alternative nested models can be compared by a likelihood-ratio test.

6.1.3 Estimation Using the sem Package in R

- The `tsls` function in the **sem** package is used to estimate structural equations by 2SLS.
  - The function works much like the `lm` function for fitting linear models by OLS, except that instrumental variables are specified in the `instruments` argument as a "one-sided" formula.
  - For example, to fit the first equation in the Duncan, Haller, and Portes model, we would specify something like

```
eqn.1 <- tsls(ROccAsp ~ RIQ + RSES + FOccAsp,
   instruments= ~ RIQ + RSES + FSES + FIQ, data=DHP)
summary(eqn.1)
```

  - This assumes that we have Duncan, Haller, and Portes's data in the data frame `DHP`, which is not the case.
- `tsls` can also perform weighted 2SLS estimation.

- The `sem` function may be used to fit a wide variety of models — including observed-variable nonrecursive models — by FIML.
- The "data" for the model may be specified either in the form of a covariance matrix (or raw-moment matrix) or as case-by-variable data in the form of an R data frame; in either case, the first argument to `sem` is a description of the model to be fit.
- For moment-matrix input, there are three required arguments:
  - `model`: A coded formulation of the model, described below.
  - `S`: The covariance matrix (or raw-moment matrix) among the observed variables in the model; may be in upper- or lower-triangular form as well as the full, symmetric matrix.
  - `N`: The number of observations on which the moment matrix is based.
  - In addition, for an observed-variable model, the argument `fixed.x` should be set to the names of the exogenous variables in the model.

- If the original data set is available it is generally advantageous to use it; for example, it is then possible to obtain robust estimates of coefficient standard errors. For data-set input, there are two required arguments:
  - `model`: As before.
  - `data`: An R data frame containing the data from which the covariance or raw moment matrix of the observed variables is computed.
  - In addition to `fixed.x`, there are two other arguments that are often useful:
    - `formula`: A one-sided R "model formula" to be applied to `data` to produce a numeric data matrix from which moments are computed; the default is `~.` .
    - `raw`: If `TRUE` (the default depends upon context but is typically `FALSE`), a raw-moment matrix is used rather than a covariance matrix, permitting the estimation of regression intercepts.
  - Additional arguments are available, e.g., to use alternative estimation criteria.

- Internally, `sem` represents the model using a format called the "recticular-action model" (or RAM), which stems from an approach, due originally to McArdle, to specifying and estimating SEMs.
- The RAM model can be specified directly using the `specifyModel` function in the **sem** package, which returns a model-specification object to be used as the first argument to `sem`:
  - Each structural coefficient of the model is represented as a directed arrow `->`.
  - Each error variance and covariance is represented as a bidirectional arrow, `<->`, linking an endogenous variables to itself or two endogenous variables, though `specifyModel` will by default supply error variances automatically for the endogenous variables in the model if these aren't given explicitly.

- To write out the model in the form required by `specifyModel`, it helps to redraw the path diagram, as in Figure 10 for the Duncan, Haller, and Portes model.
  - Then the model can be encoded as follows, specifying each arrow, and giving a name to and start-value for the corresponding parameter (`NA` = let the program compute the start-value):

```
model.DHP.1 <- specifyModel()
    RIQ     ->  ROccAsp, gamma51,  NA
    RSES    ->  ROccAsp, gamma52,  NA
    FSES    ->  FOccAsp, gamma63,  NA
    FIQ     ->  FOccAsp, gamma64,  NA
    FOccAsp ->  ROccAsp, beta56,   NA
    ROccAsp ->  FOccAsp, beta65,   NA
    ROccAsp <-> ROccAsp, sigma77,  NA
    FOccAsp <-> FOccAsp, sigma88,  NA
    ROccAsp <-> FOccAsp, sigma78,  NA
```


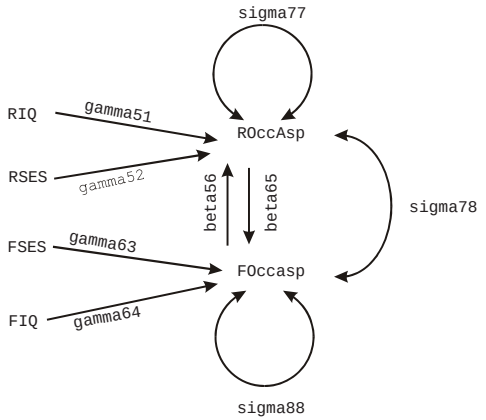

Figure 10. Modified path diagram for the Duncan, Haller, and Portes model, omitting covariances among exogenous variables, and showing error variances and covariances as double arrows attached to the endogenous variables.

- As mentioned, the error-variance parameters need not be given directly, and one can also omit the `NA`s for the start values, and so a more compact equivalent specification would be

```
model.DHP.1 <- specifyModel()
    RIQ     ->  ROccAsp, gamma51
    RSES    ->  ROccAsp, gamma52
    FSES    ->  FOccAsp, gamma63
    FIQ     ->  FOccAsp, gamma64
    FOccAsp ->  ROccAsp, beta56
    ROccAsp ->  FOccAsp, beta65
    ROccAsp <-> FOccAsp, sigma78
```


- The `specifyEquations` function is often a more convenient and compact way to specify a structural equation model; for the current example:

```
model.DHP.1 <- specifyEquations()
ROccAsp = gamma51*RIQ + gamma52*RSES + beta56*FOccAsp
FOccAsp = gamma64*FIQ + gamma63*FSES + beta65*ROccAsp
C(ROccAsp, FOccAsp) = sigma78
```

  - Each term on the RHS of a structural equation is given in the form `coefficient*explanatoryVariable`.
  - Error covariances are specified using `C()`.
  - Error variances can be specified similarly using `V()`, but this is unnecessary here since `specifyEquations` supplies them by default.

- Parameter start values can optionally be given in parentheses after the parameter name; e.g., `beta56(0.5)*FOccAsp`.

- As was common when SEMs were first introduced to sociologists, Duncan, Haller, and Porter estimated their model for standardized variables.
  - That is, the covariance matrix among the observed variables is a correlation matrix.
  - The arguments for using standardized variables in a SEM are no more compelling than in a regression model.
    - In particular, it makes no sense to standardize dummy regressors, for example.

- FIML estimates and standard errors for the Duncan, Haller, and Portes model are as follows:

Parameter	*Estimate*	*Standard Error*
γ_{51}	0.237	0.055
γ_{52}	0.176	0.046
β_{56}	0.398	0.105
γ_{63}	0.219	0.046
γ_{64}	0.311	0.058
β_{65}	0.422	0.134
σ^2_7	0.793	0.074
σ^2_8	0.717	0.088
σ_{78}	−0.495	0.139

  - The ratio of each estimate to its standard error is a Wald statistic for testing the null hypothesis that the corresponding parameter is 0, distributed asymptotically as a standard normal variable under the hypothesis.

- Note the large (and highly statistically significant) negative estimated error covariance, corresponding to an error correlation of

$$r_{78} = \frac{-0.495}{\sqrt{0.793 \times 0.717}} = -.657$$

  - I find this value implausible (a *positive* correlation would make more sense), casting doubt on the adequacy of the model.

6.2 Estimation of Recursive and Block-Recursive Models

- Because all of the explanatory variables in a structural equation of a recursive model are uncorrelated with the error, the equation can be consistently estimated by OLS.
  - For a recursive model, the OLS, 2SLS, and FIML estimates coincide.
- Estimation of a block-recursive model is essentially the same as of a nonrecursive model:
  - All variables in prior blocks are available for use as IVs in formulating 2SLS estimates.
  - FIML estimates reflect the restrictions placed on the disturbance covariances.

7. Latent Variables, Measurement Errors, and Multiple Indicators‡

- The purpose of this section is to use simple examples to explore the consequences of measurement error for the estimation of SEMs.
- I will show:
  - when and how measurement error affects the usual estimators of structural parameters;
  - how measurement errors can be taken into account in the process of estimation;
  - how multiple indicators of latent variables can be incorporated into a model.
- Then, in the next section, I will introduce and examine general structural-equation models that include these features.

‡ As time permits.

7.1 Example 1: A Nonrecursive Model With Measurement Error in the Endogenous Variables

- Consider the model displayed in the path diagram in Figure 11.
- The path diagram uses the following conventions:
  - Greek letters represent unobservables, including latent variables, structural errors, measurement errors, covariances, and structural parameters.
  - Roman letters represent observable variables.
  - Latent variables are enclosed in circles (or, more generally, ellipses), observed variables in squares (more generally, rectangles).
  - All variables are expressed as deviations from their expectations.

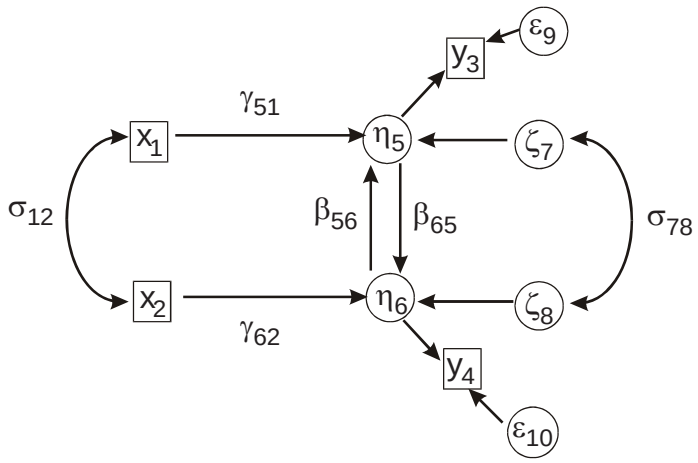

Figure 11. A nonrecursive model with measurement error in the endogenous variables.

xs	observable exogenous variables
ys	observable fallible indictors of latent endogenous variables
ηs ("eta")	latent endogenous variables
ζs ("zeta")	structural disturbances
εs ("epsilon")	measurement errors in endogenous indicators
γs, βs ("gamma", "beta")	structural parameters
σs ("sigma")	covariances

- The model consists of two sets of equations:

(a) The *structural submodel*:

$$\eta_5 = \gamma_{51}x_1 + \beta_{56}\eta_6 + \zeta_7$$
$$\eta_6 = \gamma_{62}x_2 + \beta_{65}\eta_5 + \zeta_8$$

(b) The *measurement submodel*:

$$y_3 = \eta_5 + \varepsilon_9$$
$$y_4 = \eta_6 + \varepsilon_{10}$$

- I make the usual assumptions about the behaviour of the structural disturbances — e.g., that the ζs are independent of the xs.
- I also assume "well behaved" measurement errors:
  - Each ε has an expectation of 0.
  - Each ε is independent of all other variables in the model (except the indicator to which it is attached).
- One way of approaching a latent-variable model is by substituting observable quantities for latent variables.
  - For example, working with the first structural equation:

$$\eta_5 = \gamma_{51}x_1 + \beta_{56}\eta_6 + \zeta_7$$
$$y_3 - \varepsilon_9 = \gamma_{51}x_1 + \beta_{56}(y_4 - \varepsilon_{10}) + \zeta_7$$
$$y_3 = \gamma_{51}x_1 + \beta_{56}y_4 + \zeta_7'$$

where the *composite error*, ζ_7', is

$$\zeta_7' = \zeta_7 + \varepsilon_9 - \beta_{56}\varepsilon_{10}$$

  - Because the exogenous variables x_1 and x_2 are independent of all components of the composite error, they still can be employed in the usual manner as IVs to estimate γ_{51} and β_{56}.
- Consequently, introducing measurement error into the endogenous variables of a nonrecursive model doesn't compromise our usual estimators.
  - Measurement error in an endogenous variable is not wholly benign: It does increase the size of the error variance, and thus decreases the precision of estimation.

7.2 Example 2: Measurement Error in an Exogenous Variable

- Now examine the path diagram in Figure 12.
- Some additional notation:

xs (here)	observable exogenous variable or fallible indicator of latent exogenous variable
ξ ("xi")	latent exogenous variable
δ ("delta")	measurement error in exogenous indicator

- The structural and measurement submodels are as follows:
  - structural submodel:

$$y_4 = \gamma_{46}\xi_6 + \gamma_{42}x_2 + \zeta_7$$
$$y_5 = \gamma_{53}x_3 + \beta_{54}y_4 + \zeta_8$$

  - measurement submodel:

$$x_1 = \xi_6 + \delta_9$$

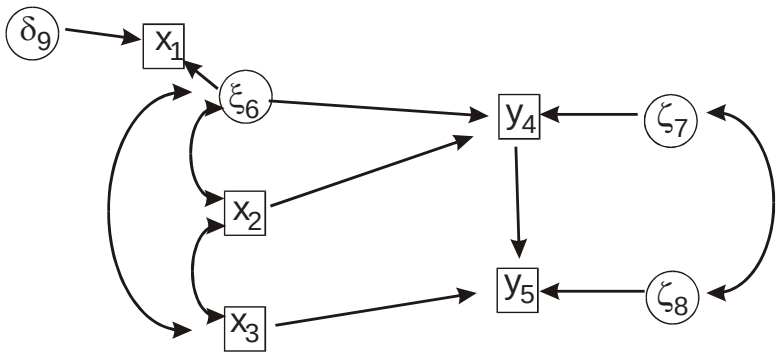

Figure 12. A structural-equation model with measurement error in an exogenous variable.

- As in the preceding example, I'll substitute for the latent variable in the first structural equation:

$$\begin{aligned} y_4 &= \gamma_{46}(x_1 - \delta_9) + \gamma_{42}x_2 + \zeta_7 \\ &= \gamma_{46}x_1 + \gamma_{42}x_2 + \zeta_7' \end{aligned}$$

where

$$\zeta_7' = \zeta_7 - \gamma_{46}\delta_9$$

is the composite error.

- If x_1 were measured without error, then we would estimate the first structural equation by OLS regression — i.e., using x_1 and x_2 as IVs.
  - Here, however, x_1 is not eligible as an IV since it is correlated with δ_9, which is a component of the composite error ζ_7'.
  - Nevertheless, to see what happens, let us multiply the rewritten structural equation in turn by x_1 and x_2 and take expectations:

$$\begin{aligned} \sigma_{14} &= \gamma_{46}\sigma_1^2 + \gamma_{42}\sigma_{12} - \gamma_{46}\sigma_9^2 \\ \sigma_{24} &= \gamma_{46}\sigma_{12} + \gamma_{42}\sigma_2^2 \end{aligned}$$

    – Notice that if x_1 is measured *without* error, then the measurement-error variance σ_9^2 is 0, and the term $-\gamma_{46}\sigma_9^2$ disappears.
  - Solving these equations for γ_{46} and γ_{42} produces

$$\begin{aligned} \gamma_{46} &= \frac{\sigma_{14}\sigma_2^2 - \sigma_{12}\sigma_{24}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2 - \sigma_9^2\sigma_2^2} \\ \gamma_{42} &= \frac{\sigma_1^2\sigma_{24} - \sigma_{12}\sigma_{14}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} - \frac{\gamma_{46}\sigma_{12}\sigma_9^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \end{aligned}$$

- Now suppose that we make the mistake of assuming that x_1 is measured without error and perform OLS estimation.
  - The OLS estimator of γ_{46} "really" estimates

$$\gamma_{46}' = \frac{\sigma_{14}\sigma_2^2 - \sigma_{12}\sigma_{24}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

  - The denominator of the equation for γ_{46} is positive, and the term $-\sigma_9^2\sigma_2^2$ in this denominator is negative, so $|\gamma_{46}'| < |\gamma_{46}|$.
    – That is, the OLS estimator of γ_{46} is biased towards zero (or *attenuated*).

- Similarly, the OLS estimator of γ_{42} really estimates

$$\begin{aligned}\gamma'_{42} &= \frac{\sigma_1^2\sigma_{24} - \sigma_{12}\sigma_{14}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ &= \gamma_{42} + \frac{\gamma_{46}\sigma_{12}\sigma_9^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ &= \gamma_{42} + \text{ bias}\end{aligned}$$

  where the bias is 0 if
  – ξ_6 does not affect y_4 (i.e., $\gamma_{46} = 0$); or
  – ξ_6 and x_2 are uncorrelated (and hence $\sigma_{12} = 0$); or
  – there is no measurement error in x_1 after all ($\sigma_9^2 = 0$).
- Otherwise, the bias can be either positive or negative; towards 0 or away from it.

- Looked at slightly differently, as the measurement error variance in x_1 grows larger (i.e., as $\sigma_9^2 \to \infty$),

$$\gamma'_{42} \to \frac{\sigma_{24}}{\sigma_2^2}$$

  - This is the population slope for the *simple* linear regression of y_4 on x_2 alone.
  - That is, when the measurement-error component of x_1 gets large, it comes an ineffective control variable as well as an ineffective explanatory variable.
- Although we cannot legitimately estimate the first structural equation by OLS regression of y_4 on x_1 and x_2, the equation is identified because both x_2 and x_3 are eligible IVs:
  - Both of these variables are uncorrelated with the composite error ζ'_7.

- It is also possible to estimate the measurement-error variance σ_9^2 and the *true-score variance* σ_6^2:
  - Squaring the measurement submodel and taking expectations produces

$$\begin{aligned}E\left(x_1^2\right) &= E[(\xi_6 + \delta_9)^2] \\ \sigma_1^2 &= \sigma_6^2 + \sigma_9^2\end{aligned}$$

    because ξ_6 and δ_9 are uncorrelated [eliminating the cross-product $E(\xi_6\delta_9)$].
  - From our earlier work,

$$\sigma_{14} = \gamma_{46}\sigma_1^2 + \gamma_{42}\sigma_{12} - \gamma_{46}\sigma_9^2$$

    – Solving for σ_9^2,

$$\sigma_9^2 = \frac{\gamma_{46}\sigma_1^2 + \gamma_{42}\sigma_{12} - \sigma_{14}}{\gamma_{46}}$$

    and so

$$\sigma_6^2 = \sigma_1^2 - \sigma_9^2$$

  – In all instances, consistent estimates are obtained by substituting observed sample variances and covariances for the corresponding population quantities.
  – the proportion of the variance of x_1 that is true-score variance is called the *reliability* of x_1; that is,

$$\text{reliability}(x_1) = \frac{\sigma_6^2}{\sigma_1^2} = \frac{\sigma_6^2}{\sigma_6^2 + \sigma_9^2}$$

  – The reliability of an indicator is also interpretable as the squared correlation between the indicator and the latent variable that it measures.
- The second structural equation of this model, for y_5, presents no difficulties because x_1, x_2, and x_3 are all uncorrelated with the structural error ζ_8 and hence are eligible IVs.

7.3 Example 3: Multiple Indicators of a Latent Variable

- Figure 13 shows the path diagram for a model that includes two different indicators x_1 and x_2 of a latent exogenous variable ξ_6.
- The structural and measurement submodels of this model are as follows;
  - Structural submodel:

$$\begin{aligned} y_4 &= \gamma_{46}\xi_6 + \beta_{45}y_5 + \zeta_7 \\ y_5 &= \gamma_{53}x_3 + \beta_{54}y_4 + \zeta_8 \end{aligned}$$

  - Measurement submodel:

$$\begin{aligned} x_1 &= \xi_6 + \delta_9 \\ x_2 &= \lambda\xi_6 + \delta_{10} \end{aligned}$$

  - Further notation:

    λ ("lambda") regression coefficient relating an indicator to a latent variable (also called a *factor loading*)

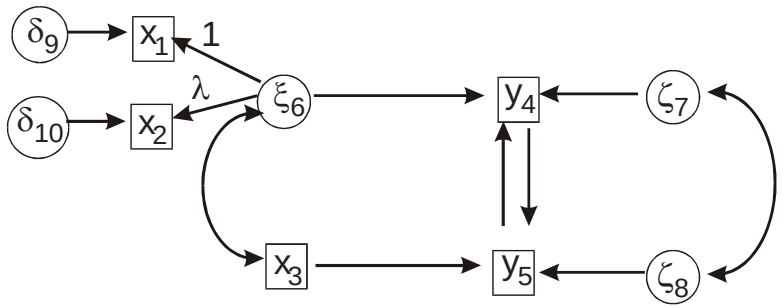

Figure 13. A model with multiple indicators of a latent variable.

- Note that one of the λs has been set to 1 to fix the scale of ξ_6.
  - That is, the scale of ξ_6 is the same as that of the *reference indicator* x_1.
  - Alternatively, the variance of the latent variable ξ_6 could be set to 1 (i.e., standardizing ξ_6).
  - Without this kind of restriction, the model is not identified.
  - This sort of scale-setting restriction is called a *normalization*.

- Once again, I will analyze the first structural equation by substituting for the latent variable ξ_6, but now that can be done in two ways:

1. using the equation for x_1,

$$\begin{aligned} y_4 &= \gamma_{46}(x_1 - \delta_9) + \beta_{45}y_5 + \zeta_7 \\ &= \gamma_{46}x_1 + \beta_{45}y_5 + \zeta_7' \end{aligned}$$

   where

$$\zeta_7' = \zeta_7 - \gamma_{46}\delta_9$$

2. using the equation for x_2,

$$\begin{aligned} y_4 &= \gamma_{46}\left(\frac{x_2}{\lambda} - \frac{\delta_{10}}{\lambda}\right) + \beta_{45}y_5 + \zeta_7 \\ &= \frac{\gamma_{46}}{\lambda}x_2 + \beta_{45}y_5 + \zeta_7'' \end{aligned}$$

   where

$$\zeta_7'' = \zeta_7 - \frac{\gamma_{46}}{\lambda}\delta_{10}$$

- Next, multiply each of these equations by x_3 and take expectations:

$$\sigma_{34} = \gamma_{46}\sigma_{13} + \beta_{45}\sigma_{35}$$
$$\sigma_{34} = \frac{\gamma_{46}}{\lambda}\sigma_{23} + \beta_{45}\sigma_{35}$$

  - These equations imply that

$$\lambda = \frac{\sigma_{23}}{\sigma_{13}}$$

- Alternative expressions for λ may be obtained by taking expectations of the two equations with the endogenous variables, y_4 and y_5, producing

$$\lambda = \frac{\sigma_{24}}{\sigma_{14}}$$

and

$$\lambda = \frac{\sigma_{25}}{\sigma_{15}}$$

  - Thus, the factor loading λ is overidentified.

  - It seems odd to use the endogenous variables y_4 and y_5 as instruments, but doing so works because they are uncorrelated with the measurement errors δ_9 and δ_{10} (and covariances involving the structural error ζ_7 cancel).

- Now apply x_2 to the first equation and x_1 to the second equation, obtaining

$$\sigma_{24} = \gamma_{46}\sigma_{12} + \beta_{45}\sigma_{25}$$
$$\sigma_{14} = \frac{\gamma_{46}}{\lambda}\sigma_{12} + \beta_{45}\sigma_{15}$$

  because x_2 is uncorrelated with ζ_7' and x_1 is uncorrelated with ζ_7''.
  - We already know λ and so these two equations can be solved for γ_{46} and β_{45}.
  - Moreover, because there is more than one way of calculating (and hence of estimating) λ, the parameters γ_{46} and β_{45} are also overidentified.

- In this model, if there were only *one* fallible indicator of ξ_6, the model would be underidentified.

8. General Structural Equation Models ("LISREL" Models)

- We now have the essential building blocks of general structural-equation models with latent variables, measurement errors, and multiple indicators, often called "LISREL" models.
  - *LISREL* is an acronym for *LI*near *S*tructural *REL*ations.
  - This model was introduced by Karl Jöreskog and his coworkers; Jöreskog and Sörbom are also responsible for the (once) widely used LISREL computer program.
- There are other formulations of general structural equation models that are equivalent to the LISREL model.

8.1 Formulation of the LISREL Model

- Several types of variables appears in LISREL models, each represented as a vector:

$\underset{(n\times 1)}{\boldsymbol{\xi}}$	("xi")	latent exogenous variables
$\underset{(q\times 1)}{\mathbf{x}}$		indicators of latent exogenous variables
$\underset{(q\times 1)}{\boldsymbol{\delta}}$	("delta")	measurement errors in the xs
$\underset{(m\times 1)}{\boldsymbol{\eta}}$	("eta")	latent endogenous variables
$\underset{(p\times 1)}{\mathbf{y}}$		indicators of latent endogenous variables
$\underset{(p\times 1)}{\boldsymbol{\varepsilon}}$	("epsilon")	measurement errors in the ys
$\underset{(m\times 1)}{\boldsymbol{\zeta}}$	("zeta")	structural disturbances

- The model also incorporates several matrices of regression coefficients:

$\underset{(m\times m)}{\mathbf{B}}$	("Beta")	structural coefficients relating ηs (latent endogenous variables) to each other
$\underset{(m\times n)}{\boldsymbol{\Gamma}}$	("Gamma")	structural coefficients relating ηs to ξs (latent endogenous to exogenous variables)
$\underset{(q\times n)}{\boldsymbol{\Lambda}_x}$	("Lambda-x")	factor loadings relating xs to ξs (indicators to latent exogenous variables)
$\underset{(p\times m)}{\boldsymbol{\Lambda}_y}$	("Lambda-y")	factor loadings relating ys to ηs (indicators to latent endogenous variables)

- Finally, there are four parameter matrices containing variances and covariances:

$\underset{(m\times m)}{\boldsymbol{\Psi}}$	("Psi")	variances and covariances of the ζs (structural disturbances)
$\underset{(q\times q)}{\boldsymbol{\Theta}_\delta}$	("Theta-delta")	variances and covariances of the δs (measurement errors in exogenous indicators)
$\underset{(p\times p)}{\boldsymbol{\Theta}_\varepsilon}$	("Theta-epsilon")	variances and covariances of the εs (measurement errors in endogenous indicators)
$\underset{(n\times n)}{\Phi}$	("Phi")	variances and covariances of the ξs (latent exogenous variables)

- The LISREL model consists of structural and measurement submodels.
  - The structural submodel is similar to the observed-variable structural-equation model in matrix form (for the ith of N observations):
$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i$$
    - Notice that the structural-coefficient matrices appear on the right-hand side of the model.
    - In this form of the model, $\mathbf{B}$ has 0s down the main diagonal.
  - The measurement submodel consists of two matrix equations, for the indicators of the latent exogenous and endogenous variables:
$$\begin{aligned}\mathbf{x}_i &= \boldsymbol{\Lambda}_x\boldsymbol{\xi}_i + \boldsymbol{\delta}_i\\ \mathbf{y}_i &= \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i\end{aligned}$$
    - Each column of the $\boldsymbol{\Lambda}$ matrices generally contains an entry that is set to 1, fixing the scale of the corresponding latent variable.
    - Alternatively, the variances of exogenous latent variables in Φ might be fixed, typically to 1.

8.2 Assumptions of the LISREL Model

- The measurement errors, $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$,
  - have expectations of $\mathbf{0}$;
  - are each multivariately-normally distributed;
  - are independent of each other;
  - are independent of the latent exogenous variables ($\boldsymbol{\xi}$s), latent endogenous variables ($\boldsymbol{\eta}$s), and structural disturbances ($\boldsymbol{\zeta}$s).
- The N observations are independently sampled.
- The latent exogenous variables, $\boldsymbol{\xi}$, are multivariate normal.
  - This assumption is unnecessary for exogenous variables that are measured without error.

- The structural disturbances, $\boldsymbol{\zeta}$,
  - have expectation $\mathbf{0}$;
  - are multivariately-normally distributed;
  - are independent of the latent exogenous variables ($\boldsymbol{\xi}$s).
- Under these assumptions, the observable indicators, $\mathbf{x}$ and $\mathbf{y}$, have a multivariate-normal distribution.

$$\begin{bmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{bmatrix} \sim N_{q+p}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ represents the population covariance matrix of the indicators.

8.3 Estimation of the LISREL Model

- The variances and covariances of the observed variables ($\boldsymbol{\Sigma}$) are functions of the parameters of the LISREL model ($\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}_x, \boldsymbol{\Lambda}_y, \boldsymbol{\Psi}, \boldsymbol{\Theta}_\delta, \boldsymbol{\Theta}_\varepsilon$, and Φ).
  - In any particular model, there will be restrictions on many of the elements of the parameter matrices.
    - Most commonly, these restrictions are exclusions: certain parameters are prespecified to be 0.
    - As I have noted, the $\boldsymbol{\Lambda}$ matrices (or the Φ matrix) must contain normalizing restrictions to set the metrics of the latent variables.
  - If the restrictions on the model are sufficient to identify it, then MLEs of the parameters can be found.

- The log-likelihood under the model is

$$\begin{aligned} &\log_e L(\mathbf{B},\boldsymbol{\Gamma}, \boldsymbol{\Lambda}_x, \boldsymbol{\Lambda}_y,\boldsymbol{\Psi}, \boldsymbol{\Theta}_\delta, \boldsymbol{\Theta}_\varepsilon, \Phi) \\ &= -\frac{N(p+q)}{2}\log_e 2\pi - \frac{N}{2}\left[\log_e \det \boldsymbol{\Sigma} + \text{trace}(\mathbf{S}\boldsymbol{\Sigma}^{-1})\right] \end{aligned}$$

  where
  - $\boldsymbol{\Sigma}$ is the covariance matrix among the observed variables that is implied by the parameters of the model.
  - $\mathbf{S}$ is the sample covariance matrix among the observed variables.
- This log-likelihood can be thought of as a measure of the proximity of $\boldsymbol{\Sigma}$ and $\mathbf{S}$, so the MLEs of the parameters are selected to make the two covariance matrices as close as possible.
- There are also other estimation criteria.

- The relationship between $\boldsymbol{\Sigma}$ and the parameters is as follows:

$$\underset{(q+p\times q+p)}{\boldsymbol{\Sigma}} = \begin{bmatrix} \underset{(q\times q)}{\boldsymbol{\Sigma}_{xx}} & \underset{(q\times p)}{\boldsymbol{\Sigma}_{xy}} \\ \underset{(p\times q)}{\boldsymbol{\Sigma}_{yx}} & \underset{(p\times p)}{\boldsymbol{\Sigma}_{yy}} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{xx} &= \boldsymbol{\Lambda}_x \Phi \boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \\
\boldsymbol{\Sigma}_{yy} &= \boldsymbol{\Lambda}_y \left[(\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Gamma} \Phi \boldsymbol{\Gamma}' (\mathbf{I} - \mathbf{B})'^{-1} + (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Psi} (\mathbf{I} - \mathbf{B})'^{-1} \right] \boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\varepsilon \\
\boldsymbol{\Sigma}_{xy} &= \boldsymbol{\Sigma}_{yx}' = \boldsymbol{\Lambda}_x \Phi \boldsymbol{\Gamma}' (\mathbf{I} - \mathbf{B})'^{-1} \boldsymbol{\Lambda}_y'
\end{aligned}$$

- As is generally the case in maximum-likelihood estimation:
  - the asymptotic standard errors for the parameter estimates may be obtained from the square-roots of the diagonal entries of the information matrix;
  - alternative nested models can be compared by a likelihood-ratio test.
  - In particular, the overidentifying restrictions on an overidentified model can be tested by comparing the maximized log-likelihood under the model with the log-likelihood of a just-identified model, which necessarily perfectly reproduces the observed sample covariances, $\mathbf{S}$.
    - The log-likelihood for a just-identified model is

$$\log_e L_1 = -\frac{N(p+q)}{2} \log_e 2\pi - \frac{N}{2} \left[\log_e \det \mathbf{S} + p + q \right]$$

    - Denoting the maximized log-likelihood for the overidentified model as $\log_e L_0$, the likelihood-ratio test statistic is, as usual, twice the difference in the log-likelihoods for the two models:

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

    - Under the hypothesis that the overidentified model is correct, this statistic is distributed as chi-square, with degrees of freedom equal to the degree of overidentification of the model, that is, the difference between the number of variances and covariances among the observed variables in the model, which is

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

    and the number of free parameters in the model.

- One can also compute standard errors and tests that are robust with respect to non-normality.

8.4 Identification of LISREL Models[§]

- Identification of models with latent variables is a complex problem without a simple general solution.
- A global necessary condition for identification is that the number of free parameters in the model can be no larger than the number of variances and covariances among observed variables,

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

  - Unlike the order condition for observed-variable nonrecursive models, this condition is insufficiently restrictive to give us any confidence that a model that meets the condition is identified.
  - That is, it is easy to meet this condition and still have an underidentified model.

[§] As time permits.

- A useful rule that sometimes helps is that a model is identified if:
  (a) all of the measurement errors in the model are uncorrelated with one-another;
  (b) there are at least two unique indicators for each latent variable, or if there is only one indicator for a latent variable, it is measured without error;
  (c) the structural submodel would be identified were it an observed-variable model.
- The likelihood function for an underidentified model flattens out at the maximum, and consequently
  - the maximum isn't unique; and
  - the information matrix is singular
- Computer programs for structural-equation modelling can usually detect an attempt to estimate an underidentified model, or will produce output that is obviously incorrect.

8.5 Examples

8.5.1 A Latent-Variable Model for the Peer-Influences Data

- Figure 14 shows a latent-variable model for Duncan, Haller, and Portes's peer-influences data.

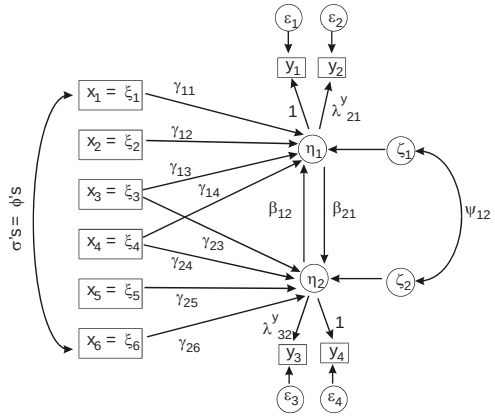

Figure 14. Latent-variable model for the peer-influences data.

- The variables in the model are as follows:

x_1 (ξ_1)	respondent's parents' aspirations
x_2 (ξ_2)	respondent's family IQ
x_3 (ξ_3)	respondent's SES
x_4 (ξ_4)	best friend's SES
x_5 (ξ_5)	best friend's family IQ
x_6 (ξ_6)	best friend's parents' aspirations
y_1	respondent's occupational aspiration
y_2	respondent's educational aspiration
y_3	best friend's educational aspiration
y_4	best friend's occupational aspiration
η_1	respondent's general aspirations
η_2	best friend's general aspirations

- In this model, the exogenous variables each have a single indicator specified to be measured without error, while the latent endogenous variables each have two fallible indicators.

- The structural and measurement submodels are as follows:
  - Structural submodel:

$$\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}0 & \beta_{12}\\ \beta_{21} & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & 0 & 0\\ 0 & 0 & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26}\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\\ \xi_4\\ \xi_5\\ \xi_6\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix}$$

$$\Psi = \text{Var}\left(\begin{bmatrix}\zeta_1\\ \zeta_2\end{bmatrix}\right) = \begin{bmatrix}\psi_1^2 & \psi_{12}\\ \psi_{12} & \psi_2^2\end{bmatrix} \text{ (note: symmetric)}$$

- Measurement submodel:

$$\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6\end{bmatrix} = \begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\\ \xi_4\\ \xi_5\\ \xi_6\end{bmatrix}; \text{ i.e., } \boldsymbol{\Lambda}_x = \mathbf{I}_6, \boldsymbol{\Theta}_\delta = \underset{(6\times 6)}{\mathbf{0}}, \text{ and } \Phi = \underset{(6\times 6)}{\boldsymbol{\Sigma}_{xx}}$$

$$\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\end{bmatrix} = \begin{bmatrix}1 & 0\\ \lambda_{21}^y & 0\\ 0 & \lambda_{32}^y\\ 0 & 1\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\end{bmatrix}, \text{ with } \boldsymbol{\Theta}_\varepsilon = \text{diag}(\theta_{11}^\varepsilon, \theta_{22}^\varepsilon, \theta_{33}^\varepsilon, \theta_{44}^\varepsilon)$$

- We can specify this model for `sem` as follows:

```
model.dhp.2 <- specifyEquations(covs="RGenAsp, FGenAsp")
RGenAsp = gam11*RParAsp + gam12*RIQ + gam13*RSES
              + gam14*FSES + beta12*FGenAsp
FGenAsp = gam23*RSES + gam24*FSES + gam25*FIQ
              + gam26*FParAsp + beta21*RGenAsp
ROccAsp = 1*RGenAsp
REdAsp = lam21*RGenAsp
FOccAsp = 1*FGenAsp
FEdAsp = lam42*FGenAsp
```

  - `sem` assumes that variables that do not appear in the data (here, `RGenAsp` and `FGenAsp`) are latent variables.

- The argument `covs="RGenAsp, FGenAsp"` to `specifyEquations` includes error variance and covariance parameters for the two latent endogenous variables, and is an alternative to using the `C()` and `V()` operators.
- Because `RParAsp`, `RIQ`, `RSES`, `FSES`, `FIQ`, and `FParAsp` are directly observed exogenous variables, these should be specified in the `fixed.x` argument to `sem`.

- Maximum-likelihood estimates of the parameters of the model and their standard errors:

Parameter	*Estimate*	*Std. Error*	*Parameter*	*Estimate*	*Std. Error*
γ_{11}	0.161	0.038	λ^y_{21}	1.063	0.092
γ_{12}	0.250	0.045	λ^y_{42}	0.930	0.071
γ_{13}	0.218	0.043	ψ^2_1	0.281	0.046
γ_{14}	0.072	0.050	ψ^2_2	0.264	0.045
γ_{23}	0.062	0.052	ψ_{12}	−0.023	0.052
γ_{24}	0.229	0.044	θ^ε_{11}	0.412	0.052
γ_{25}	0.349	0.045	θ^ε_{22}	0.336	0.053
γ_{26}	0.159	0.040	θ^ε_{33}	0.311	0.047
β_{12}	0.184	0.096	θ^ε_{44}	0.405	0.047
β_{21}	0.235	0.120			

- With the exception of $\widehat{\gamma}_{14}$ and $\widehat{\gamma}_{23}$, the direct effect of each boy's SES on the other's aspirations, all of the coefficients of the exogenous variables are statistically significant.
- The reciprocal paths, $\widehat{\beta}_{12}$ and $\widehat{\beta}_{21}$, have respective p-values just smaller than and just larger than $.05$ for a two-sided test, but a one-sided test would be appropriate here anyway.
- The negative covariance between the structural disturbances, $\widehat{\psi}_{12} = -0.023$, is now close to 0 and non-significant, though a positive value would be even more plausible.

- Because the indicator variables are standardized in this model, the measurement-error variances represent the proportion of variance of each indicator due to measurement error, and the complements of the measurement-error variances are the reliabilities of the indicators.
  - For example, the estimated reliability of y_1 (the respondent's reported occupational aspiration) as an indicator η_1 (his general aspirations) is $1 - 0.412 = .588$.
  - Further details are in the computer examples.

8.5.2 A Confirmatory-Factor-Analysis Model

- The LISREL model is very general, and special cases of it correspond to a variety of statistical models.
- For example, if there are only exogenous latent variables and their indicators, the LISREL models specializes to the *confirmatory-factor-analysis* (*CFA*) model, which seeks to account for the correlational structure of a set of observed variables in terms of a smaller number of factors.
- The path diagram for an illustrative CFA model appears in Figure 15.
  - The data for this example are taken from Harman's classic factor-analysis text.
  - Harman attributes the data to Holzinger, an important figure in the development of factor analysis (and intelligence testing).

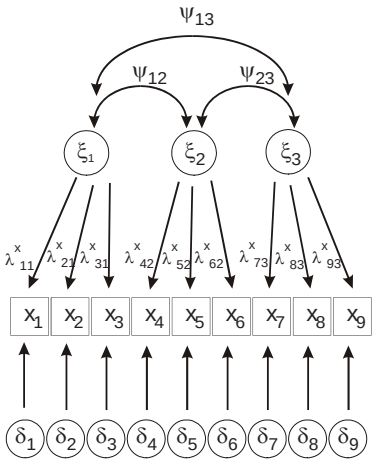

Figure 15. A confirmatory-factor-analysis model for three factors underlying nine psychological tests.

- The first three tests (Word Meaning, Sentence Completion, and Odd Words) are meant to tap a verbal factor; the next three (Mixed Arithmetic, Remainders, Missing Numbers) an arithmetic factor, and the last three (Gloves, Boots, Hatchets) a spatial-relations factor.
- The model permits the three factors to be correlated with one-another.
- The normalizations employed in this model set the variances of the factors to 1; the covariances of the factors are then the factor intercorrelations.

▶ This model can be conveniently specified using the `cfa` function in the **sem** package:

```
model.Holzinger.2 <- cfa(reference.indicators=FALSE)
Verbal: Word.meaning, Sentence.completion, Odd.words
Arithmetic: Mixed.arithmetic, Remainders,
        Missing.numbers
Spatial: Gloves, Boots, Hatchets
```

- Each factor is given a name, followed by a colon and the names of the observed variables loading on that factor.
- The argument `reference.indicators=FALSE` sets the factor variances to 1 rather than the loading of the first indicator for each factor to 1.

- By default, the factors are assumed to be correlated, and their pairwise correlations (or covariances) are free parameters to be estimated from the data; including the argument `covs=NULL` would specify uncorrelated (“orthogonal”) factors.

▶ Estimates for this model, and for an alternative CFA model specifying uncorrelated factors, are given in the computer examples.

9. Other Capabilities of the sem Package¶

- Robust standard errors and test statistics.
- FIML estimates in the presence of missing data.
- Multiple imputation of missing data, using the `mi` package.
- Ordinal indicators and bootstrapped standard errors.
- Multiple-group SEMs.
- Alternative estimation criteria (objective functions).
- Alternative optimizers.

¶ Many to be illustrated as time permits.