The R Statistical Computing Environment Basics and Beyond Structural Equation Models with the **sem** package

John Fox

McMaster University

ICPSR/Berkeley 2016

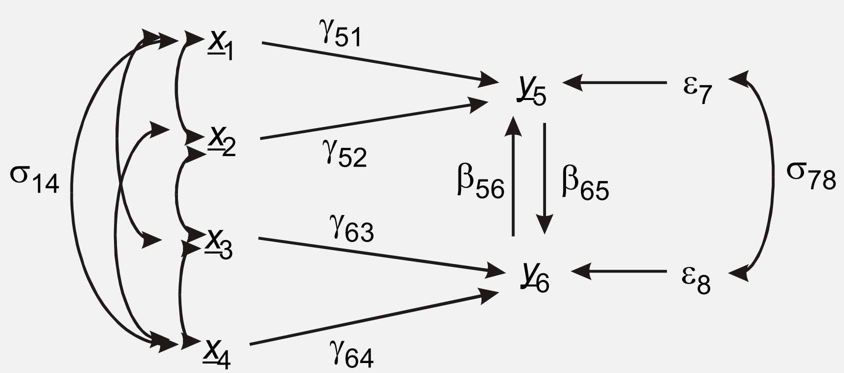
Nonrecursive Model for Peer-Influences Data

Variables in the Model

- A nonrecursive model, from Duncan, Haller, and Portes's (1968) study of peer influences on the aspirations of high-school students, appears in the following figure.
- Variables:
  - 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.

Nonrecursive Model for Peer-Influences Data

Path Diagram



Nonrecursive Model for Peer-Influences Data

Conventions 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.

Nonrecursive Model for Peer-Influences Data
Structural Equations

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

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

Nonrecursive Model for Peer-Influences Data
Structural Equations

- I'll usually simplify the structural equations by
  1. suppressing the subscript i for observation;
  2. expressing all xs and ys 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:

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

Nonrecursive Model for Peer-Influences Data
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.

Nonrecursive Model for Peer-Influences Data
Estimation Using the **sem** Package in R

- 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.

Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- 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.

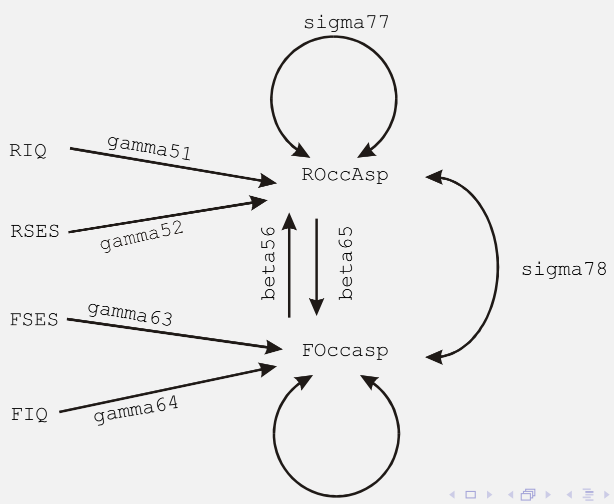
Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- 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 the following figure.

Nonrecursive Model for Peer-Influences Data

Modified path diagram omitting covariances among exogenous variables, and showing error variances and covariances as double arrows attached to the endogenous variables.



Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- 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
```

Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- As mentioned, the error-variance parameters need not be given directly, and one can also omit the NAs 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
```


Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- 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()`.

Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- 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`.
- Fixed parameters can be specified using numeric constants; e.g. (not pertaining to the Duncan, Haller, and Portes data), `1*age`.

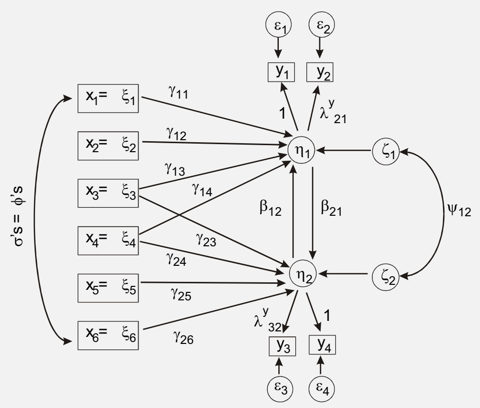
Nonrecursive Model for Peer-Influences Data

Estimation Using the **sem** Package in R

- 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.

A Latent-Variable Model for the Peer-Influences Data

Path Diagram



A Latent-Variable Model for the Peer-Influences Data

Variables in the Model

x_1	respondent's parents' aspirations
x_2	respondent's family IQ
x_3	respondent's SES
x_4	best friend's SES
x_5	best friend's family IQ
x_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 are specified to be measured without error, while the latent endogenous variables each have two fallible indicators.

A Latent-Variable Model for the Peer-Influences Data

Structural Equations

- Measurement submodel:

$$
\begin{aligned}
y_1 &= \eta_1 + \varepsilon_1 \\
y_2 &= \lambda_{21}\eta_1 + \varepsilon_2 \\
y_3 &= \lambda_{31}\eta_2 + \varepsilon_3 \\
y_4 &= \eta_2 + \varepsilon_4
\end{aligned}
$$

- Structural submodel:

$$
\begin{aligned}
\eta_1 &= \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3 + \beta_{12}\eta_2 + \zeta_1 \\
\eta_2 &= \gamma_{24}x_4 + \gamma_{25}x_5 + \gamma_{26}x_6 + \beta_{21}\eta_1 + \zeta_2
\end{aligned}
$$

A Latent-Variable Model for the Peer-Influences Data

Coding the Model for sem

- 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
```

A Latent-Variable Model for the Peer-Influences Data

Coding the Model for sem

- `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`.

A Confirmatory-Factor-Analysis Model

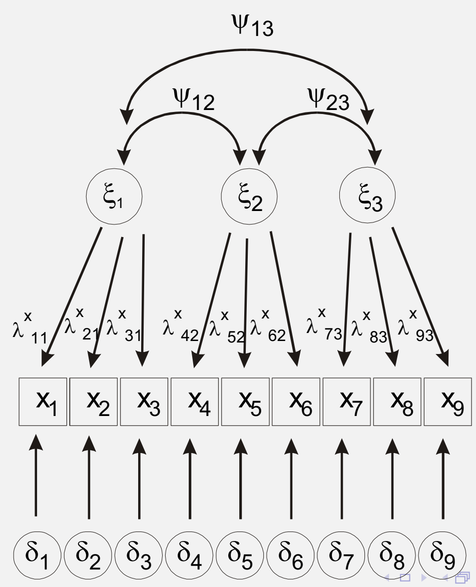
- The latent-variable structural equation model is very general, and special cases of it correspond to a variety of statistical models.
- For example, if there are only exogenous latent variables and their indicators, the model specializes to the *confirmatory-factor-analysis* (*CFA*) model, which seeks to account for the covariational structure of a set of observed variables in terms of a smaller number of factors.

A Confirmatory-Factor-Analysis Model

- 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).
- 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.

A Confirmatory-Factor-Analysis Model

Path Diagram



A Confirmatory-Factor-Analysis Model

Coding the Model using cfa

- 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; including the argument covs=NULL would specify uncorrelated ("orthogonal") factors.

Additional Capabilities of the **sem** Package and Other SEM software in R

- Additional features 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.
- Other R packages for structural equation modeling:
  - **lavaan**, general structural equation models
  - **OpenMx**, general structural equation models
  - **systemfit**, observed variables structural equation models