STATISTICAL HORIZONS

Structural Equation Modeling

Paul D. Allison, Ph.D.

Upcoming Seminar:
July 16-20, 2018, Chicago, Illinois

Structural Equation Modeling

Paul D. Allison, Instructor

www.StatisticalHorizons.com

Structural Equation Models

The classic SEM model includes many common linear models used in the behavioral sciences:

- Multiple regression
- ANOVA
- Path analysis
- Multivariate ANOVA and regression
- Factor analysis
- Canonical correlation
- Non-recursive simultaneous equations
- Seemingly unrelated regressions
- Dynamic panel data models

What is SEM good for?

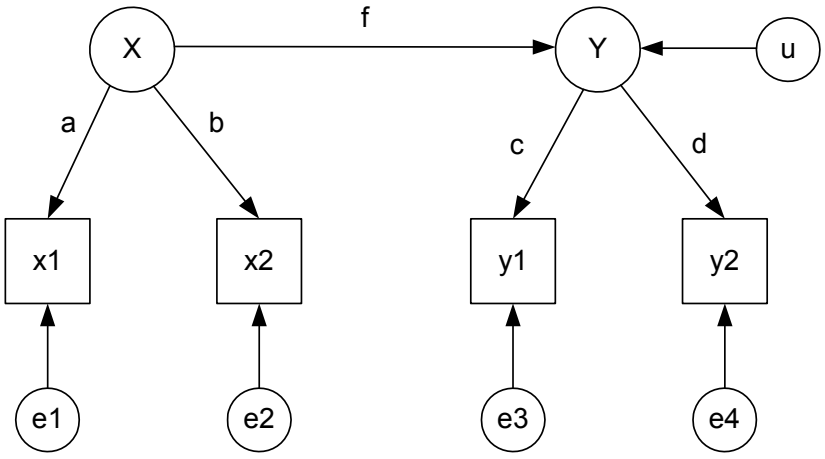
- Modeling complex causal mechanisms.
- Studying mediation (direct and indirect effects).
- Correcting for measurement error in predictor variables.
- Avoiding multicollinearity for predictors variables that are measuring the same thing.
- Analysis with instrumental variables.
- Modeling reciprocal relationships (2-way causation).
- Handling missing data (by maximum likelihood).
- Scale construction and development.
- Analyzing longitudinal data.
- Providing a very general modeling framework to handle all sorts of different problems in a unified way.

SEM

Convergence of psychometrics and econometrics

- Simultaneous equation models, possibly with reciprocal (nonrecursive) relationships
- Latent (unobserved) variables with multiple indicators.
- This course emphasizes models with latent variables. For example:

Preview: A Latent Variable SEM Model



X and Y are unobserved variables, x1, x2, y1, and y2 are observed indicators, e1-e4 and u are random errors.
a, b, c, d, and f are correlation coefficients.

Latent Variable Model (cont.)

- If we know the six correlations among the observed variables, simple hand calculations can produce estimates of *a* through *f*. We can also test the fit of the model.
- Why is it desirable to estimate models like this?
  - Most variables are measured with at least some error.
  - In a regression model, measurement error in independent variables can produce severe bias in coefficient estimates.
  - We can correct this bias if we have multiple indicators for variables with measurement error.
  - Multiple indicators can also yield more powerful hypothesis tests.

Cautions

- Although SEM's can be very useful, the methodology is often used badly and indiscriminately.
  - Often applied to data where it's inappropriate.
  - Can sometimes obscure rather than illuminate.
  - Easy to get sucked into overly complex modeling.

Outline

1. Introduction to SEM
2. Linear regression with missing data
3. Path analysis of observed variables
4. Direct and indirect effects
5. Identification problem in nonrecursive models
6. Reliability: parallel and tau-equivalent measures
7. Multiple indicators of latent variables
8. Confirmatory factor analysis
9. Goodness of fit measures
10. Structural relations among latent variables
11. Alternative estimation methods.
12. Multiple group analysis
13. Models for ordinal and nominal data
14. Longitudinal Data Analysis

Software for SEMs

LISREL – Karl Jöreskog and Dag Sörbom

EQS – Peter Bentler

PROC CALIS (SAS) – W. Hartmann, Yiu-Fai Yung

Amos – James Arbuckle

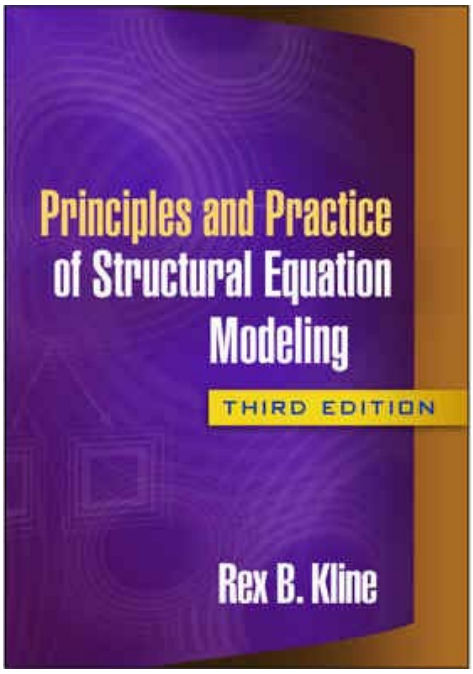
Mplus – Bengt Muthén

sem, gsem (Stata)

Packages for R:

- **OpenMX** – Michael Neale
- **sem** – John Fox
- **lavaan (R)** – Yves Rosseel

Favorite Textbook



Linear Regression in SEM

The standard linear regression model is just a special case of SEM:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

We make the usual assumptions about ε:

- uncorrelated with the x's.
- mean of 0
- homoskedastic (variance is constant)
- normally distributed.

By default, all SEM programs do maximum likelihood (ML) estimation. Under these assumptions, ML is equivalent to ordinary least squares (OLS).

Why do it in SEM? Because SEM can handle missing data by maximum likelihood—one of the best methods available.

GSS2014 Example

Data from the 2014 General Social Survey (GSS). There were a total of 2538 respondents. Here are the variables that we will use, along with their ranges and the number of cases with data missing:

AGE	Age of respondent (18-89), 9 cases missing
ATTEND	Frequency of attendance at religious services (0-8), 13 cases missing
CHILDS	Number of children (0-8), 8 cases missing
EDUC	Highest year of school completed (0-20), 1 case missing
FEMALE	1=female, 0=male
HEALTH	Condition of health (1 excellent – 4 poor), 828 cases missing; 824 of these were not asked the question
INCOME	Total family income (in thousands of dollars), 224 cases missing
MARRIED	1=married, 0=unmarried, 4 cases missing
PAEDUC	Father's highest year school completed, father (0 – 20), 653 cases missing
PARTYID	Political party identification (1 strong democrat – 6 strong republican); 88 cases missing
POLVIEWS	Think of self as liberal or conservative (1 liberal – 7 conservative) 89 cases missing
PROCHOICE	Scale of support for abortion rights (1 – 6), 1033 cases missing; 824 of these were not asked the question **(dependent variable)**
WHITE	1=white race, 0= non-white

Regression with Mplus

```
DATA:
  FILE = c:\data\gss2014.csv;
VARIABLE:
  NAMES = age attend childs educ health income paeduc partyid
    polviews female married white prochoice; MISSING = .;
  USEVARIABLES = age attend childs educ health income paeduc
     female married white prochoice;
MODEL:
  prochoice ON age attend childs educ health income paeduc
     female married white;
```

Mplus only reads text files, without any variable names.

My convention: All upper case words are Mplus key words; all lower case words are variable names, parameter names, or data set names that you choose. (Mplus is not case sensitive).

Mplus doesn't have a default missing data code, so we have to assign it with the MISSING option.

USEVARIABLES is necessary to limit the variables to those actually used in the model.

Mplus Output

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PROCHOIC ON				
AGE	0.013	0.004	3.457	0.001
ATTEND	-0.292	0.021	-13.932	0.000
CHILDS	-0.087	0.042	-2.048	0.041
EDUC	0.132	0.023	5.682	0.000
HEALTH	-0.139	0.072	-1.926	0.054
INCOME	0.004	0.001	3.809	0.000
PAEDUC	0.035	0.016	2.183	0.029
FEMALE	-0.048	0.114	-0.422	0.673
MARRIED	-0.378	0.127	-2.983	0.003
WHITE	-0.496	0.145	-3.416	0.001
Intercepts				
PROCHOICE	2.665	0.420	6.337	0.000

1480 cases are lost because of missing data.

Linear Regression with Stata

```
use "c:\data\gss2014.dta"
sem prochoice <- age attend childs educ health income
   paeduc female married white
```

```
 |                            OIM
                |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+------------------------------------------------------------------
Structural      |
  prochoice <-  |
           age |   .0132813   .0038416     3.46   0.001     .0057518    .0208108
        attend |  -.2923748   .0209863   -13.93   0.000    -.3335072   -.2512424
        childs |   -.087023   .0424846    -2.05   0.041    -.1702913   -.0037547
          educ |    .132245   .0232747     5.68   0.000     .0866275    .1778626
        health |  -.1393584   .0723665    -1.93   0.054    -.2811941    .0024773
        income |   .0043811   .0011504     3.81   0.000     .0021264    .0066357
        paeduc |    .034699    .015891     2.18   0.029     .0035532    .0658448
        female |  -.0479459   .1135087    -0.42   0.673     -.270419    .1745271
       married |  -.3777956   .1266759    -2.98   0.003    -.6260758   -.1295153
         white |  -.4960648    .145236    -3.42   0.001    -.7807221   -.2114074
         _cons |   2.664199   .4204945     6.34   0.000     1.840045    3.488353
```


Linear Regression with SAS

```
PROC CALIS DATA=my.gss2014;
  PATH prochoice <- age attend childs educ health
      income paeduc female married white; RUN;
```

Like Mplus, SAS is not case sensitive, but I put SAS language in upper case. PATH is one of 7 "languages" in PROC CALIS for specifying SEMs.

PATH List

Path			Parameter	Estimate	Standard Error	t Value	Pr > |t|
prochoice	<===	age	_Parm01	0.01328	0.00384	3.4556	0.0005
prochoice	<===	attend	_Parm02	-0.29237	0.02100	-13.9251	<.0001
prochoice	<===	childs	_Parm03	-0.08702	0.04250	-2.0474	0.0406
prochoice	<===	educ	_Parm04	0.13225	0.02329	5.6792	<.0001
prochoice	<===	health	_Parm05	-0.13936	0.07240	-1.9248	0.0543
prochoice	<===	income	_Parm06	0.00438	0.00115	3.8066	0.0001
prochoice	<===	paeduc	_Parm07	0.03470	0.01590	2.1825	0.0291
prochoice	<===	female	_Parm08	-0.04795	0.11356	-0.4222	0.6729
prochoice	<===	married	_Parm09	-0.37780	0.12674	-2.9810	0.0029
prochoice	<===	white	_Parm10	-0.49606	0.14530	-3.4140	0.0006

Linear Regression with lavaan

```
gssdata<-read.table("c:/data/gss2014_names.txt", header=T)
gssmod <- ' prochoice ~ age+attend+childs+educ+health
                +income+paeduc+female+married+white '
gssfit <- sem(gssmod, data=gssdata)
summary(gssfit)
```

For R, missing data must be coded as NA. ~ means "is regressed on". For this program to work, the lavaan package must be installed into R.

```
prochoice ~       Estimate  Std.Err  Z-value  P(>|z|)
  age                0.013    0.004    3.457    0.001
  attend            -0.292    0.021  -13.932    0.000
  childs            -0.087    0.042   -2.048    0.041
  educ               0.132    0.023    5.682    0.000
  health            -0.139    0.072   -1.926    0.054
  income             0.004    0.001    3.808    0.000
  paeduc             0.035    0.016    2.184    0.029
  female            -0.048    0.114   -0.422    0.673
  married           -0.378    0.127   -2.982    0.003
  white             -0.496    0.145   -3.416    0.001
```


FIML for Missing Data

Full information maximum likelihood (FIML) is one of the best (and easiest) methods for dealing with missing data. It has several advantages over multiple imputation:

- Because multiple imputation introduces random variation into the imputation process, you get a different result every time you use it. FIML always produces the same result.
- Multiple imputation is a rather complex and "messy" method that requires a lot of attention to detail. With FIML, you just specify one option.
- Multiple imputation requires two separate models, an imputation model and an analysis model. If they're not compatible you may get incorrect results. With FIML, there's only one model, so no danger of incompatibility.
- FIML is fully efficient, in the statistical sense. Multiple imputation is close to fully efficient, but doesn't quite get there.

Further Reading

Allison, Paul D. (2003) "Missing data techniques for structural equation models." *Journal of Abnormal Psychology* 112: 545-557.

Download at
http://www.statisticalhorizons.com/resources/articles

Assumptions

Both FIML and MI assume that the data are missing at random (MAR): roughly, the probability that a variable has missing data does not depend on the value of that variable, once other variables are controlled.

- This would be violated, for example, if people with higher incomes were less likely to report their incomes.

FIML (and some versions of multiple imputation) assumes that variables with missing data have a multivariate normal distribution.

FIML Theory 1

The first step in ML is to construct the likelihood function, which expresses the probability of the data as a function of the unknown parameters.

Suppose that we have *n* independent observations (*i* =1,…, *n*) on *k* variables (y_{i1}, y_{i2},…, y_{ik}) and no missing data. The likelihood function is then

$$L(\theta) = \prod_{i=1}^{n} f_i(y_{i1}, y_{i2}, \ldots, y_{ik}; \theta)$$

where $f_i(.)$ is the joint probability (or probability density) function for observation *i*, and θ is a set of parameters to be estimated. To get the ML estimates, we find the values of θ that make *L* as large as possible.

Now suppose that for a particular observation *i*, the first two variables, y_1 and y_2, have missing data that satisfy the MAR assumption. The contribution to the likelihood function for that observation is just the probability of observing the remaining variables, y_{i3} through y_{ik}. How do we get that?

- If y_1 and y_2 are discrete, we sum the joint probability over all possible values of the two variables with missing data:

$$f_i^*(y_{i3}, \ldots, y_{ik}; \theta) = \sum_{y_1} \sum_{y_2} f_i(y_{i1}, \ldots, y_{ik}; \theta)$$

FIML Theory 2

If the missing variables are continuous, we use integrals in place of summations:

$$f_i^*(y_{i3}, \ldots, y_{ik}; \theta) = \int_{y_1} \int_{y_2} f_i(y_{i1}, y_{i2}, \ldots y_{ik}) dy_2 dy_1$$

The overall likelihood is just the product of the likelihoods for all the observations. For example, if there are *m* observations with complete data and *n-m* observations with data missing on y_1 and y_2, the likelihood function for the full data set becomes

$$L(\theta) = \prod_{i=1}^{m} f_i(y_{i1}, y_{i2}, \ldots, y_{ik}; \theta) \prod_{i=m+1}^{n} f_i^*(y_{i3}, , \ldots, y_{ik}; \theta)$$

where observations are ordered such that the first *m* have no missing data and the last *n-m* have missing data. This likelihood can then be maximized to get ML estimates of θ.

FIML Theory 3

In the case of linear models, we invoke the multivariate normal assumption. When no data are missing, the likelihood function is

$$L(\theta) = \prod_i f(\mathbf{y}_i \mid \boldsymbol{\mu}(\theta), \boldsymbol{\Sigma}(\theta))$$

where $\mathbf{y}_i$ is a vector of all the observed variables and the density function is given by

$$f(\mathbf{y}) = \frac{\exp[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})]}{(2\pi)^{k/2} \mid \boldsymbol{\Sigma} \mid^{1/2}}$$

When data are missing (at random), the likelihood becomes

$$L(\theta) = \prod_i f(\mathbf{y}_i \mid \mu_i(\theta), \Sigma_i(\theta))$$

- If data are missing for individual *i,* then $\mathbf{y}_i$ deletes the missing values, $\boldsymbol{\mu}_i$ deletes the corresponding means, and $\boldsymbol{\Sigma}_i$ deletes the corresponding rows and columns.
- This likelihood can be maximized by conventional methods, e.g., the Newton-Raphson algorithm or the EM algorithm.

FIML in SAS

```
PROC CALIS DATA=my.gss2014 METHOD=FIML;
  PATH prochoice <- age attend childs educ health
      income paeduc female married white; RUN;
```

Path			Parameter	Estimate	Standard Error	t Value	Pr > |t|
prochoice	<===	age	_Parm01	0.01330	0.00325	4.0915	<.0001
prochoice	<===	attend	_Parm02	-0.26510	0.01779	-14.8973	<.0001
prochoice	<===	childs	_Parm03	-0.07401	0.03372	-2.1950	0.0282
prochoice	<===	educ	_Parm04	0.13026	0.02000	6.5133	<.0001
prochoice	<===	health	_Parm05	-0.03465	0.06027	-0.5749	0.5654
prochoice	<===	income	_Parm06	0.00499	0.00106	4.7204	<.0001
prochoice	<===	paeduc	_Parm07	0.04205	0.01578	2.6651	0.0077
prochoice	<===	female	_Parm08	0.03062	0.09745	0.3142	0.7534
prochoice	<===	married	_Parm09	-0.30837	0.10866	-2.8378	0.0045
prochoice	<===	white	_Parm10	-0.38116	0.11743	-3.2457	0.0012

FIML in Stata

```
use "c:\data\gss2014.dta"
sem prochoice <- age attend childs educ health  income
    paeduc female married white, method(mlmv)
```

MLMV stands for maximum likelihood treatment of missing values.

```
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  prochoice <- |
         age |   .0133017    .003251     4.09   0.000     .0069298    .0196736
      attend |  -.2650952   .0177948   -14.90   0.000    -.2999724   -.2302181
      childs |  -.0740106   .0337186    -2.19   0.028    -.1400978   -.0079234
        educ |   .1302599   .0199992     6.51   0.000     .0910622    .1694577
      health |  -.0346473   .0602697    -0.57   0.565    -.1527737    .0834792
      income |   .0049948   .0010582     4.72   0.000     .0029208    .0070687
      paeduc |   .0420557   .0157793     2.67   0.008      .011129    .0729825
      female |    .030612   .0974469     0.31   0.753    -.1603803    .2216043
     married |  -.3083652   .1086642    -2.84   0.005    -.5213432   -.0953872
       white |  -.3811773   .1174348    -3.25   0.001    -.6113454   -.1510092
       _cons |   2.078671   .3449224     6.03   0.000     1.402635    2.754706
```


FIML in lavaan

```
gssdata<-read.table("c:/data/gss2014_names.txt", header=T)
gssmod <- ' prochoice ~ age+attend+childs+educ+health
                +income+paeduc+female+married+white '
gssfit <- sem(gssmod, data=gssdata, missing='fiml')
summary(gssfit)
```

```
  Number of observations                          2538
  Number of missing patterns                        33

Regressions:
                   Estimate  Std.Err  Z-value  P(>|z|)
  prochoice ~
    age               0.013    0.003    4.093    0.000
    attend           -0.265    0.018  -14.899    0.000
    childs           -0.074    0.034   -2.196    0.028
    educ              0.130    0.020    6.517    0.000
    health           -0.035    0.060   -0.575    0.565
    income            0.005    0.001    4.721    0.000
    paeduc            0.042    0.016    2.669    0.008
    female            0.031    0.097    0.314    0.753
    married          -0.308    0.109   -2.838    0.005
    white            -0.381    0.117   -3.248    0.001
```

FIML in Mplus

```
DATA: FILE = c:\data\gss2014.csv;
VARIABLE:
  NAMES = age attend childs educ health income paeduc partyid
    polviews female married white prochoice; MISSING = .;
  USEVARIABLES = age attend childs educ health income paeduc
     female married white prochoice;
MODEL:
  prochoice ON age attend childs educ health income paeduc
     female married white;
  age attend childs educ health income paeduc
     female married white;
```

Mplus does FIML by default, but only for dependent variables. To make it work for independent variables, we must name them on a separate statement. The names refer to their variances, which tells Mplus to treat them as if they were dependent.

Mplus "Problem"

Mplus gives the same results as SAS and Stata. But it also reports the following:

```
THE MODEL ESTIMATION TERMINATED NORMALLY

THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES MAY NOT BE
TRUSTWORTHY FOR SOME PARAMETERS DUE TO A NON-POSITIVE DEFINITE
FIRST-ORDER DERIVATIVE PRODUCT MATRIX.  THIS MAY BE DUE TO THE STARTING
VALUES BUT MAY ALSO BE AN INDICATION OF MODEL NONIDENTIFICATION.  THE
CONDITION NUMBER IS      -0.957D-19.  PROBLEM INVOLVING THE FOLLOWING
PARAMETER:
     Parameter 77, WHITE
```

This is not a real problem. It happens whenever Mplus tries to estimate a variance for a dummy variable. For a dummy variable, the variance is a function of the mean. Whenever one parameter is a function of another, Mplus flags it as a possible indication of non-identification.

FIML with Auxiliary Variables

FIML can be improved by the inclusion of auxiliary variables—variables that are correlated with variables that have missing data, but are not themselves in the model.

- To include auxiliary variables, allow them to be freely correlated with all the variables in the regression model.
- An easy way to do that is to specify additional regression equations in which the auxiliary variables are dependent variables. But we must also allow the auxiliary variables to be correlated with each other.
- SEM packages vary greatly in how auxiliary variables may be included.
- For the GSS2014 data, PARTYID and POLVIEWS can be used as auxiliary variables.

Auxiliary Variables in Mplus

```
DATA: FILE = c:\data\gss2014.csv;
VARIABLE:
  NAMES = age attend childs educ health income paeduc partyid
    polviews female married white prochoice; MISSING = .;
  AUXILIARY = (M) polviews partyid;
MODEL:
  prochoice ON age attend childs educ health income paeduc
     female married white;
  age attend childs educ health income paeduc
     female married white;
```

Mplus has a special syntax for auxiliary variables.

We no longer need the USEVARIABLES option because all the variables are being used.

Results differ only slightly from model without auxiliary variables. Standard errors are slightly smaller.

Auxiliary Variables in SAS

```
PROC CALIS DATA=my.gss2014 METHOD=FIML;
  PATH prochoice <- age attend childs educ health
      income paeduc female married white,
  partyid polviews <- prochoice age attend childs educ
     health income paeduc female married white,
  partyid <-> polviews;
RUN;
```

The second "path" specifies two regressions. The sole function of these regressions is to allow the auxiliary variables to be correlated with all the others.

The third path with the <-> operator specifies the partial correlation between these the auxiliary variables.

Auxiliary Variables in Stata

```
use "c:\data\gss2014.dta"
sem (prochoice<-age attend childs educ health income
      paeduc female married white)
   (partyid polviews <- prochoice age attend childs
   educ health income paeduc female married white),
   cov(e.partyid*e.polviews) method(mlmv)
```

The second "path" specifies two regressions. The sole function of these regressions is to allow the auxiliary variables to be correlated with all the others.

The **cov** option specifies the partial correlation between these the auxiliary variables.

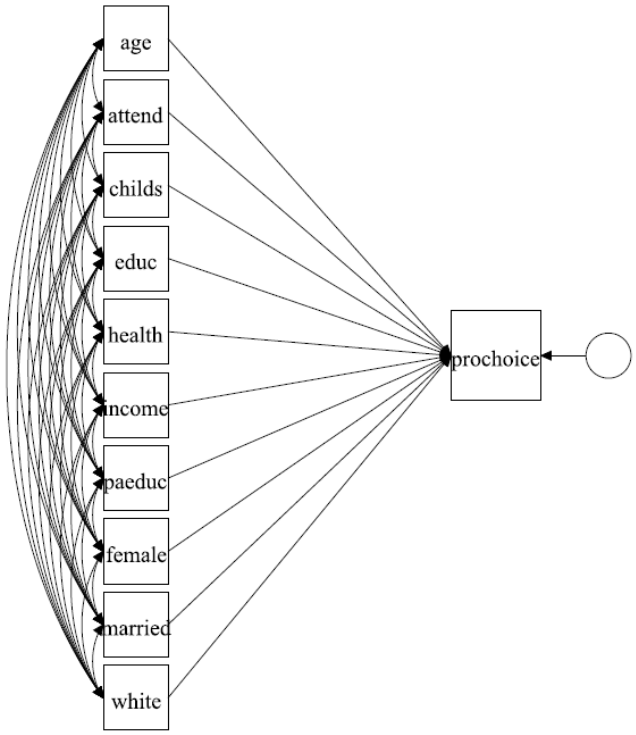
Auxiliary Variables in lavaan

```
gssdata<-read.table("c:/data/gss2014_names.txt", header=T)
gssmod <- ' prochoice ~ age+attend+childs+educ+health
                  +income+paeduc+female+married+white
   polviews+partyid ~ prochoice+age+attend+childs+educ+health
                  +income+paeduc+female+married+white
   polviews ~~ partyid '
gssfit <- sem(gssmod, data=gssdata, missing='fiml')
summary(gssfit)
```

The second equation specifies two regressions. The sole function of these regressions is to allow the auxiliary variables to be correlated with all the others.

The third path with the ~~ operator specifies the partial correlation between these the auxiliary variables.

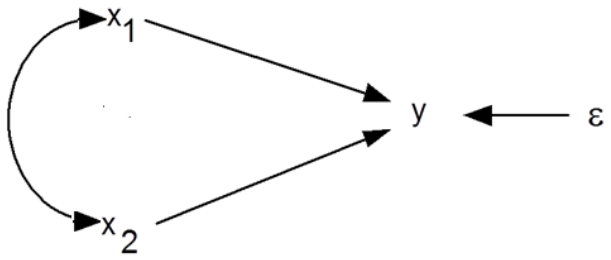
Path Diagram from Mplus



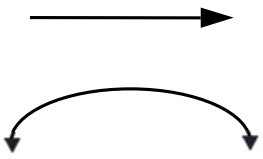
Path Analysis of Observed Variables

In the SEM literature, it's common to represent a linear model by a path diagram.

- A diagrammatic method for representing a system of linear equations. There are precise rules so that you can write down equations from looking at the diagram.
- Invented by the geneticist Sewall Wright in 1934.
- Single equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



Some Rules and Definitions



Direct causal effect

Correlation
(no causal assumptions)

Why the curved double-headed arrow in the diagram?
Because omitting it implies no correlation between x_1 and x_2.

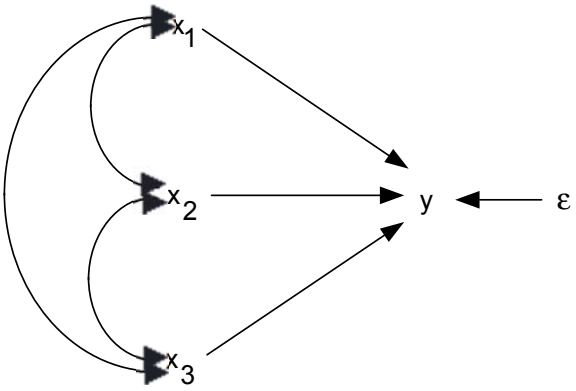
Endogenous variables: Variables caused by other variables in the system. These variables have straight arrows leading into them.

Exogenous variables: Variables not caused by others in the system. No straight arrows leading into them.

Not the same as dependent and independent because a variable that is dependent in one equation and independent in another equation is still endogenous.

Curved double-headed arrows can only link *exogenous* variables.

Three Predictor Variables



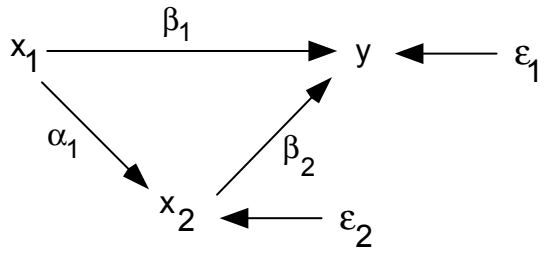
The fact that there are no curved arrows between ε and the x's implies that $\rho_{1\varepsilon} = 0$, $\rho_{2\varepsilon} = 0$, and $\rho_{3\varepsilon} = 0$. We make this assumption in the usual linear regression model.

Two-Equation System

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1$$

$$x_2 = \alpha_0 + \alpha_1 x_1 + \varepsilon_2$$

The diagram is now



Note: The diagram goes further than the equations by asserting that

$$\rho_{\varepsilon_1\varepsilon_2} = 0,\ \rho_{\varepsilon_1 x_1} = 0,\ \rho_{\varepsilon_1 x_2} = 0,\ \rho_{x_1\varepsilon_2} = 0$$

Why combine the two equations?

Answer: to get further insight into the causal process.

To make this more concrete, let's suppose that

y = income

x_1 = father's income

x_2 = years of schooling

What happens when you increase x_1 by one unit? Then y changes by β_1 units, holding x_2 constant.

This can be misleading, however, because a one-unit increase in x_1 *also* produces a change of α_1 units in x_2, which in turn produces a change in y.

Thus x_1 has both a *direct* and an *indirect* effect on y. You wouldn't notice this with a single equation.

Schooling *mediates* part of the effect of father's income on income.

Calculation of Indirect Effect

Substitute one equation into the other.

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\alpha_0 + \alpha_1 x_1 + \varepsilon_2) + \varepsilon_1$$

$$= (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)x_1 + (\varepsilon_1 + \beta_2\varepsilon_2)$$

The direct effect of x_1 is β_1 .

The indirect effect of x_1 is $\alpha_1\beta_2$.

The total effect of x_1 is $\beta_1 + \alpha_1\beta_2$

For recursive systems, indirect effects may be calculated by taking the product of coefficients along a particular path.