STATISTICAL HORIZONS

Longitudinal Data Analysis Using Structural Equation Modeling

Paul D. Allison, Ph.D.

Upcoming Seminar:

April 4-May 2, 2022, On Demand

Longitudinal Data Analysis Using SEM

Paul D. Allison, Ph.D.

Causal Inference

How do you make causal inferences with non-experimental panel data?

What's panel data?

- Data in which variables are measured at multiple points in time for the same individuals.
  - Response variable y_{it} with t = 1, 2,…, T; i = 1,…, N
  - Vector of predictor variables x_{it} .
  - Some of the predictors may vary with time, others may not.
- Assume that time points are the same for everyone in the sample.
  - Nice if they are equally spaced, but not essential.
  - We will eventually allow for drop out and other kinds of missing data.

Causal Inference

How do you demonstrate that *x* causes *y*?

- Show that *x* is correlated with *y*.
- Show that the correlation is not produced by other variables that affect both *x* and *y*.
- Show that *y* is not causing *x*.
- Show that the correlation is not due to chance alone.

Randomized experiments are great at all these things.

Panel data make it possible to

- Control for unobserved variables.
- Estimate the effect of *x* on *y*, even if *y* is also affecting *x*.

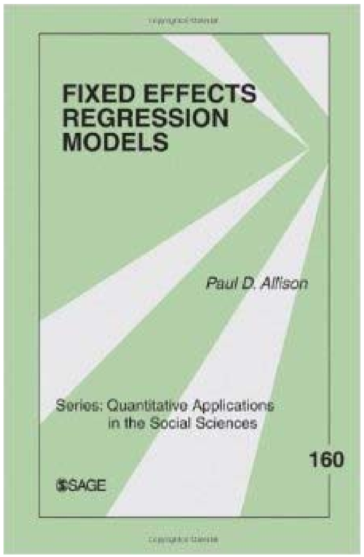
No standard methods can do both of these things simultaneously.

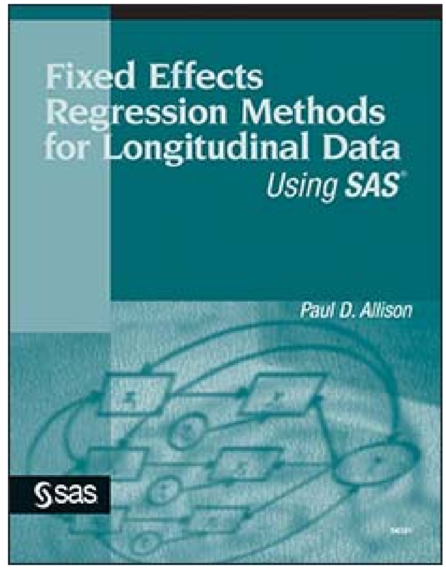
Fixed Effects Methods

To control for unobservables, we can used *fixed effects* methods

- These control for all unchanging variables whether observed or not.
- For linear models, the most common way to do fixed effects is to express all variables as deviations from individual-specific means. But there are several alternative approaches.
    - Implement FE with **xtreg** in Stata or PROC GLM in SAS
- Downsides:
    - Standard errors go up (because you're only using within individual variation).
    - Many methods don't produce estimates for time-invariant predictors.

Some References





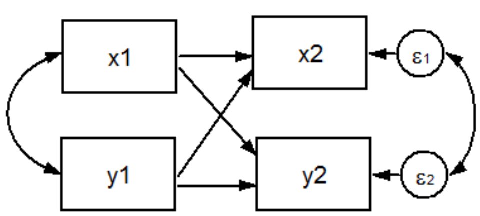
2009 2005

Cross-Lagged Linear Models

To allow for reciprocal causation, estimate 2-wave, 2-variable panel model (OD Duncan 1969) by ordinary least squares:

$$y_2 = b_0 + b_1 y_1 + b_2 x_1 + \varepsilon_2$$

$$x_2 = a_0 + a_1 y_1 + a_2 x_1 + \varepsilon_1$$



Inclusion of lagged dependent variable is intended to control for past characteristics of the individual.

Among those with the same value of y_1, b_2 is the effect of x_1 on y_2.

Our Goal

To be able to estimate models that combine fixed effects with cross-lags using structural equation modeling software. The models look like this:

Cross-lagged Effect

Fixed Effect

$$y_{it} = \mu_t + \beta_1 x_{i,t-1} + \beta_2 y_{i,t-1} + \delta_1 w_{it} + \gamma_1 z_i + \alpha_i + \varepsilon_{it}$$

$$x_{it} = \tau_t + \beta_3 x_{i,t-1} + \beta_4 y_{i,t-1} + \delta_2 w_{it} + \gamma_2 z_i + \eta_i + \upsilon_{it}$$

Cross-lagged Effect

Fixed Effect

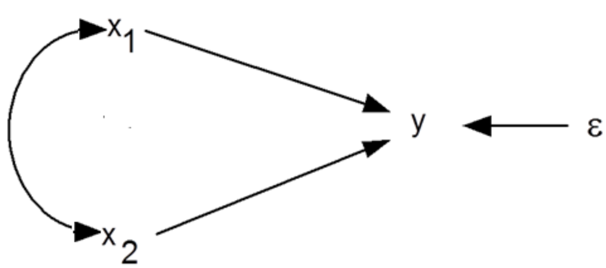
To get there, we'll

- Review models with cross-lagged effects using SEM.
- Review conventional fixed effects
- See how to do fixed effects with SEM
- Combine the two methods

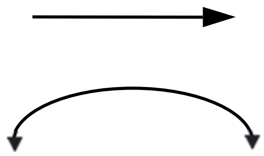
Path Analysis of Observed Variables

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.
- 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.

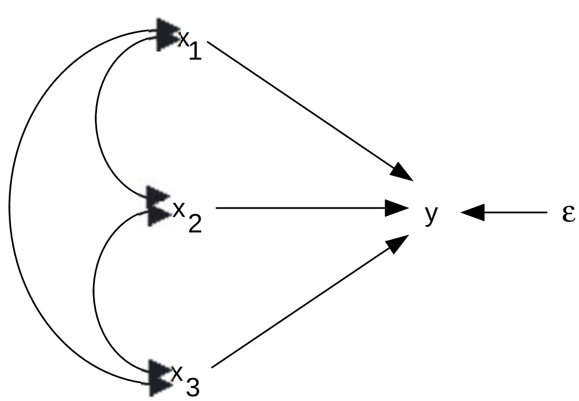
<u>Endogenous variables:</u> Variables caused by other variables in the system. These variables have straight arrows leading into them.

<u>Exogenous variables:</u> 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



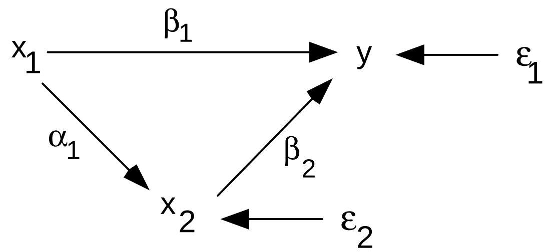
The fact that there are no curved arrows between ε and the x's implies that $\rho_{1\varepsilon} = 0$, $\rho_{2\varepsilon} = 0$, and $\rho_{3\varepsilon} = 0$. We make this assumption in the usual linear regression model.

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



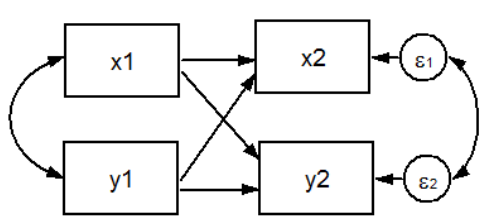
Note: The diagram goes further than the equations by asserting that

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

Cross-Lagged Linear Models

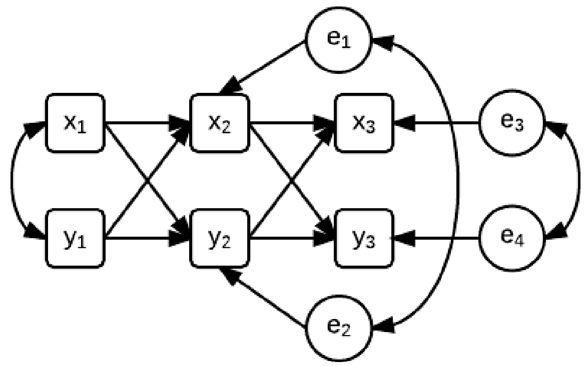
$y_2 = b_0 + b_1 y_1 + b_2 x_1 + \varepsilon_2$

$x_2 = a_0 + a_1 y_1 + a_2 x_1 + \varepsilon_1$



- This model can be estimated by ordinary least squares for each equation separately.
- Other predictors could also be included in each equation.
- Presumes no simultaneous causation.

3 Wave-2 Variable Model



- Can extend to more waves
- Each of the 4 equations could be estimated by OLS
- Can estimate simultaneously via SEM
  - Can constrain coefficients to be the same across waves.
  - Can test the overall fit of the model (OK to omit lag-2 effects?)
  - Can handle missing data by full information maximum likelihood.

NLSY Data Set

581 children interviewed in 1990, 1992, and 1994 as part of the National Longitudinal Survey of Youth (NLSY).

Time-varying variables (measured at each of the three time points):

ANTI	antisocial behavior, measured with a scale from 0 to 6.
SELF	self-esteem, measured with a scale ranging from 6 to 24.
POV	poverty status of family, coded 1 for family in poverty, otherwise 0.

Time-invariant variables:

BLACK	1 if child is black, otherwise 0
HISPANIC	1 if child is Hispanic, otherwise 0
CHILDAGE	child's age in 1990
MARRIED	1 if mother was currently married in 1990, otherwise 0
GENDER	1 if female, 0 if male
MOMAGE	mother's age at birth of child
MOMWORK	1 if mother was employed in 1990, otherwise 0

Data are in the "wide form": one record for each child , with different names for the variables at each time point, e.g., ANTI1, ANTI2 and ANTI3.

Estimating a Cross-Lagged Model

- We'll estimate the 3W-2V panel model with SEM to answer the question, does antisocial behavior affect self-esteem, or does self-esteem affect antisocial behavior?
- Other variables could be included, but we'll leave them out for simplicity.
- Cross-sectionally, these variables are significantly correlated at about -.15.
- Important to allow for correlated errors. Why? Other factors affecting both variables are not included.
- No missing data in this data set.
- We'll see how to do it with Mplus, PROC CALIS in SAS, **sem** in Stata and **lavaan** for R.

Software for SEMs

LISREL – Karl Jöreskog and Dag Sörbom

EQS – Peter Bentler

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

OpenMX (R) – Michael Neale

Amos – James Arbuckle

Mplus – Bengt Muthén

sem, gsem (Stata)

lavaan (R) – Yves Rosseel

SAS Program

```
PROC CALIS DATA=my.nlsy PSHORT;
PATH
 anti3 <- anti2 self2,
 anti2 <- anti1 self1,
 self3 <- anti2 self2,
 self2 <- anti1 self1,
 anti3 <-> self3,
 anti2 <-> self2;
RUN;
```

This option suppresses some of the voluminous output from CALIS.

A correlation between two endogenous variables is a partial correlation. That is, a correlation between their error terms.

- My convention: upper case words are part of the SAS language, lower case words are variables or data set names specific to this example. SAS is not case sensitive.
- PATH is one of 7 different "languages" for specifying SEM's.
- <- means "is regressed on". <-> means "is correlated with".

Goodness of Fit Results

Absolute Index	Fit Function	0.0827
	Chi-Square	47.9911
	Chi-Square DF	4
	Pr > Chi-Square	<.0001
	Z-Test of Wilson & Hilferty	5.7057
	Hoelter Critical N	115
	Root Mean Square Residual (RMR)	0.1956
	Standardized RMR (SRMR)	0.0388
	Goodness of Fit Index (GFI)	0.9740
Parsimony Index	Adjusted GFI (AGFI)	0.8637
	Parsimonious GFI	0.2597
	RMSEA Estimate	0.1377
	RMSEA Lower 90% Confidence Limit	0.1044
	RMSEA Upper 90% Confidence Limit	0.1739

Parameter Estimates

PATH List							
Path			Parameter	Estimate	Standard Error	t Value	Pr > |t|
anti3	<===	anti2	_Parm01	0.66063	0.03659	18.0534	<.0001
anti3	<===	self2	_Parm02	0.02148	0.01617	1.3286	0.1840
anti2	<===	anti1	_Parm03	0.67219	0.03411	19.7081	<.0001
anti2	<===	self1	_Parm04	-0.01604	0.01572	-1.0207	0.3074
self3	<===	anti2	_Parm05	0.01908	0.08200	0.2327	0.8160
self3	<===	self2	_Parm06	0.35218	0.03623	9.7209	<.0001
self2	<===	anti1	_Parm07	-0.12979	0.09474	-1.3701	0.1707
self2	<===	self1	_Parm08	0.35208	0.04366	8.0650	<.0001
anti3	<==>	self3	_Parm09	-0.62324	0.17139	-3.6363	0.0003
anti2	<==>	self2	_Parm10	-0.65746	0.16819	-3.9090	<.0001

These are unstandardized estimates.

Stata Program

```
use "C:\data\nlsy.dta", clear
sem (anti94 <- anti92 self92) ///
    (anti92 <- anti90 self90) ///
    (self94 <- anti92 self92) ///
    (self92 <- anti90 self90), ///
   cov(e.anti94*e.self94 e.anti92*e.self92)
```

- Stata is case sensitive
- <- means “is regressed on”
- e.anti94 refers to the error term for anti94
- The **cov** option allows for covariances (and therefore correlations) between pairs of variables.
- /// goes to a new a line within a single command, in a DO file.

Stata Results

```
|                            OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  anti94 <-  |
      anti92 |   .6606303   .0365617    18.07   0.000     .5889707    .7322899
      self92 |   .0214813   .0161547     1.33   0.184    -.0101814    .0531441
       _cons |   .2555298   .3492365     0.73   0.464    -.428961     .9400207
  -----------+----------------------------------------------------------------
  anti92 <-  |
      anti90 |   .6721907   .0340779    19.73   0.000     .6053992    .7389822
      self90 |  -.016042    .0157035    -1.02   0.307    -.0468202    .0147362
       _cons |   .8635114   .3296234     2.62   0.009     .2174614    1.509561
  -----------+----------------------------------------------------------------
  self92 <-  |
      anti90 |  -.1297929   .0946541    -1.37   0.170    -.3153115    .0557257
      self90 |   .3520793   .0436176     8.07   0.000     .2665903    .4375682
       _cons |   13.49922   .9155554    14.74   0.000     11.70476    15.29367
  -----------+----------------------------------------------------------------
  self94 <-  |
      anti92 |   .0190816   .0819248     0.23   0.816    -.1414881    .1796512
      self92 |   .3521843   .0361984     9.73   0.000     .2812368    .4231319
       _cons |   13.41623   .7825436    17.14   0.000     11.88247    14.94999
```


Stata Results (cont.)

```
-------------+----------------------------------------------------------------
var(e.anti94)|    1.81788   .1066577                      1.620406    2.039419
var(e.anti92)|   1.436514   .0842824                      1.280467    1.611577
var(e.self92)|   11.08263   .6502342                      9.878743    12.43324
var(e.self94)|   9.127317   .5355131                      8.135831    10.23963
-------------+----------------------------------------------------------------
cov(e.anti94,|
    e.self94)|  -.6221672   .1709519    -3.64   0.000    -.9572267   -.2871077
cov(e.anti92,|
    e.self92)|   -.656332    .167759    -3.91   0.000    -.9851335   -.3275305
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(4)   =     48.07, Prob > chi2 = 0.0000
```

The LR (likelihood ratio) test is testing the null hypothesis that all four two-period lagged paths are 0. Clearly, that must be rejected.

Mplus Program

```
DATA: FILE = c:\data\nlsy.dat;
VARIABLE: NAMES = anti90 anti92 anti94 black
 childage gender hispanic married momage momwork
 pov90 pov92 pov94 self90 self92 self94;
 USEVARIABLES = anti90 anti92 anti94 self90 self92 self94;
MODEL:
  anti94 ON anti92 self92;
  self94 ON anti92 self92;
  anti92 ON anti90 self90;
  self92 ON anti90 self90;
  anti94 WITH self94;
  anti92 WITH self92;
```

Must be a text file with no names

Necessary to restrict the variables to those actually used in the model

WITH specifies a correlation. If the two variables are endogenous, it is a correlation between their error terms, i.e., a partial correlation

Mplus is not case sensitive. But, for clarity, I capitalize words that are part of the Mplus language.

Mplus – Goodness of Fit

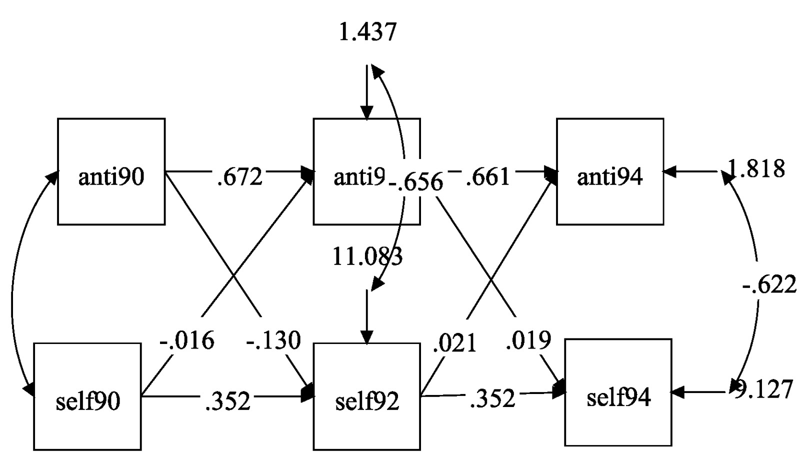
```
Chi-Square Test of Model Fit

          Value                                48.074
          Degrees of Freedom                        4
          P-Value                              0.0000

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                              0.138
          90 Percent C.I.                       0.104   0.174
          Probability RMSEA <= .05              0.000

CFI/TLI
          CFI                                   0.944
          TLI                                   0.804
```

We want the RMSEA to be <.05, definitely not above .10.
We want the CFI and TLI to be close to 1, definitely not below .90.

Mplus – Parameter Estimates

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
ANTI94 ON				
ANTI92	0.661	0.037	18.069	0.000
SELF92	0.021	0.016	1.330	0.184
SELF94 ON				
ANTI92	0.019	0.082	0.233	0.816
SELF92	0.352	0.036	9.729	0.000
ANTI92 ON				
ANTI90	0.672	0.034	19.725	0.000
SELF90	-0.016	0.016	-1.022	0.307
SELF92 ON				
ANTI90	-0.130	0.095	-1.371	0.170
SELF90	0.352	0.044	8.072	0.000
ANTI94 WITH				
SELF94	-0.622	0.171	-3.640	0.000
ANTI92 WITH				
SELF92	-0.656	0.168	-3.912	0.000

Mplus - Path Diagram



lavaan Program

```
nlsy<-read.table("C:/data/nlsy-names.txt",header=T)
nlsymod<-'
  anti94 ~ anti92 + self92
  self94 ~ anti92 + self92
  anti92 ~ anti90 + self90
  self92 ~ anti90 + self90
  anti94 ~~ self94
  anti92 ~~ self92 '
nlsyfit<-sem(nlsymod,data=nlsy)
summary(nlsyfit)
```

Slashes must be forward. This file has variables names as the first record.

Note single quotes.

~ means "is regressed on"
~~ means "is correlated with"

lavaan Results

```
Minimum Function Test Statistic                  48.074
  Degrees of freedom                                  4
  P-value (Chi-square)                            0.000

Regressions:
                   Estimate  Std.err  Z-value  P(>|z|)
  anti94 ~
    anti92            0.661    0.037   18.069    0.000
    self92            0.021    0.016    1.330    0.184
  self94 ~
    anti92            0.019    0.082    0.233    0.816
    self92            0.352    0.036    9.729    0.000
  anti92 ~
    anti90            0.672    0.034   19.725    0.000
    self90           -0.016    0.016   -1.022    0.307
  self92 ~
    anti90           -0.130    0.095   -1.371    0.170
    self90            0.352    0.044    8.072    0.000

Covariances:
  anti94 ~~
    self94           -0.622    0.171   -3.639    0.000
  anti92 ~~
    self92           -0.656    0.168   -3.912    0.000
```

Estimation & Assumptions

By default, all of these SEM packages do maximum likelihood (ML) estimation:

- Choose parameter estimates so that the probability of observing what has actually been observed is as large as possible.
- Under most conditions, ML estimators are consistent, asymptotically efficient, and asymptotically normal (if all the assumptions are met).

Assumptions:

- The specified relationships are correct.
- The endogenous variables have a multivariate normal distribution, which implies
  - All variables are normally distributed.
  - All conditional expectation functions are linear.
  - All conditional variance functions are homoscedastic.

Parameter estimates are robust to violations of multivariate normality, but chi-squares may be too large and standard errors too small.

Chi-Square Test

- If the specified model is correct, the chi-square statistic has approximately a chi-square distribution. The df is equal to the number of overidentifying restrictions (number of sample moments minus the number of parameters in the model).
- This statistic is a likelihood ratio chi-square comparing the fitted model with a saturated (just-identified) model that perfectly fits the data. If the chi-square is large and the *p*-value is small, it's an indication that the model should be rejected.
- Although this statistic is properly regarded as a test of the model, note that it is only testing the overidentifying restrictions.
- This test is sensitive to sample size. With a large sample, it may be difficult to find any parsimonious model that passes this test.