STATISTICAL
HORIZONS

Longitudinal Data Analysis Using Structural Equation Modeling

Paul Allison, Ph.D.

Upcoming Seminar:
August 17-18, 2017, Stockholm, Sweden

1 □ **Longitudinal Data Analysis Using sem**

2 □ **Causal Inference**

3 □ **Causal Inference**

4 □ **Fixed Effects Methods**

5 □ **Some References**

6 □ **Cross-Lagged Linear Models**

7 □ **Our Goal**

8 □ **Path Analysis of Observed Variables**

9 □ **Some Rules and Definitions**

10 □ **Three Predictor Variables**

11 □ **Two-Equation System**

12 □ **Cross-Lagged Linear Models**

13 □ **3 Wave-2 Variable Model**

14 □ **NLSY Data Set**

15 □ **Estimating a Cross-Lagged Model**

16 □ **Software for SEMs**

17 □ **Stata Program**

18 □ **Stata Results**

19 □ **Stata Results (cont.)**

20 □ **Path Diagram**

21 □ **Estimation & Assumptions**

22 □ **Chi-Square Test**

23 □ **Other Measures of Fit**

24 □ **Other Measures of Fit**

25 □ **Global Goodness of Fit Measures**

26 □ **Other Global Measures**

27 □ **Equality Constraints**

28 □ **Stata Program with Constraints**

29 □ **Constrained Estimates**

30 □ **Fixed Effects Analysis**

31 □ **Fixed Effects Model**

32 □ **Estimating the FE Model**

33 □ **Switching Data From Wide to Long**

34 □ **Stata for Method 2 with NLSY Data**

35 □ **Limitations of Classic FE**

36 □ **FE in SEM**

37 □ **FE with sem command**

38 □ **Sem Results**

39 □ **Sem Results (cont.)**

40 □ **Standardized Results**

41 □ **Goodness of Fit**

42 □ **Path Diagram (from Mplus)**

43 □ **Random Effects Model**

44 □ **Random vs. Fixed Effects**

45 □ **Intro to xtdpdml**

46 □ **Defaults for xtdpdml**

47 □ **FE with xtdpdml**

48 □ **xtdpdml Output**

49 □ **Time-Invariant Predictors in FE Model**

50 □ **Time-Invariant Predictors in Stata**

51 □ **Invariant Predictors with xtdpdml**

52 □ **What if Alpha and the Time-Invariant Predictors Are Allowed to Correlate?**

53 □ **Adding Lagged Predictors to FE Model**

54 □ **Issues with Lagged Predictors**

55 □ **Diagram for FE with Lagged Predictors**

56 □ **FE with Lagged Predictors in Stata**

57 □ **Same Model with xtdpdml**

58 □ **What if *y* Also Affects *x*?**

59 □ **Difference Scores in Stata**

60 □ **If *T*>3, Things Get Trickier**

61 □ **Correlations for Ousey Data**

62 □ **Model for *T*=4**

63 □ **Model Diagram**

64 □ **Stata Program for Ousey Data**

65 □ **Stata Output - GOF**

66 □ **Stata Output - Estimates**

67 □ **Stata Output – Standardized Estimates**

68 □ **Model with Only 1 Fixed Effect**

69 □ **Handling Missing Data with ML**

70 □ **Further Reading**

71 □ **FIML in Stata**

72 □ **Alternative: One Direction at a Time**

73 □ **Unidirectional Model**

74 □ **Tricking Stata**

75 □ **Results**

76 □ **Alternative Trick**

77 □ **Unidirectional with xtdpdml**

78 □ **Econometric Approach: Arellano-Bond**

79 □ **xtabond in Stata**

80 □ **xtabond for victimization**

81 □ **Problems with Arellano-Bond**

82 □ **Superiority of ML-SEM**

83 □ **New Example: Hip Data**

84 □ **Conventional Analysis in Long Form**

85 □ **Fixed Effects for Hip Data**

86 □ **Fixed Effects in Wide Form – Stata**

87 □ **Improving the Fit**

88 □ **Relaxing the Constraint in Stata**

89 □ **Relaxing the Constraint with xtdpdml**

90 □ **xtdpdml Results**

91 □ **Goodness of Fit**

92 □ **Likelihood Ratio Comparison**

93 □ **Allow for Reverse Causation**

94 □ **Reverse Causation in Stata**

95 □ **Likelihood Ratio Comparison**

96 □ **Lagging CESD and SRH**

97 □ **Lagged Variables in Stata**

98 □ **Lagged Variables with xtdpdml**

99 □ **GOF Results**

100 □ **Parameter Estimates**

101 □ **Are Lagged Effects Equal Over Time?**

102 □ **Lagged Estimates**

103 □ **Compare the Two Models**

104 □ **Lagged Effect of CESD on SRH**

105 □ **Getting the Lags Right**

106 □ **Wrong Lags**

107 □ **What To Do About Lags?**

108 □ **Victim/Offender Study**

109 □ **Stata Results**

110 □ **Linear Effect of Time**

111 □ **Linear Time Effect in Stata**

112 □ **Stata Results**

113 □ **Random (Latent) Growth Curve Model**

114 □ **Random Time Effect in Stata**

115 □ **Selected Results**

116 □ **Compare Random and
Fixed Slope Models**

117 □ **Quadratic Time Effect**

118 □ **When Must Time Be Continuous?**

119 □ **Higher Level Clustering**

120 □ **Fixed Effects With Clustering**

121 □ **Robust Standard Errors**

122 □ **Stata with Robust Standard Errors**

123 □ **Discrete Dependent Variables**

124 □ **Teen Poverty Data**

125 □ **Fixed Effects Logit in Stata**

126 □ **Results for FE Logit Model**

127 □ **Allow Effects of Alpha to Vary**

128 □ **Unconstrained Results**

129 □ **Compare the Two Models**

130 □ **Allow Lagged Effects of POV**

131 □ **Results With Lagged POV**

132 □ **Compare Lagged with Unlagged Model**

133 □ **A Lagged Effect of POV on HOURS**

134 □ **gsem Code for Reciprocal Effect**

135 □ **Results With Reciprocal Effect**

136 □ **Let’s Lag Hours**

137 □ **Results with Lagged Hours**

138 □ **Models for Count Data**

139 □ **Poisson & Negative Binomial Models**

140 □ **Conventional FE for Patent Data**

141 □ **Conventional FE in Stata**

142 □ **Stata FE nbreg Results**

143 □ **FE NB with SEM in Stata**

144 □ **Stata Results**

145 □ **Allow Coefficients for Alpha to Vary**

146 □ **Allow Patents to Affect Later R&D**

147 □ **Allow Patents to Affect Later Patents**

148 □ **Include Science and LogSize**

149 □ **Summary**

Longitudinal Data Analysis Using **sem**

Paul D. Allison, Ph.D.

February 2016

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*.

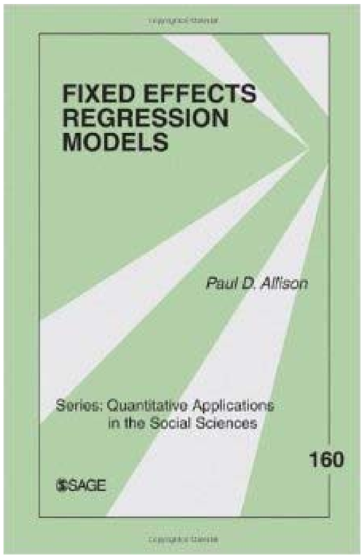
No standard methods can do both of these things simultaneously.

Fixed Effects Methods

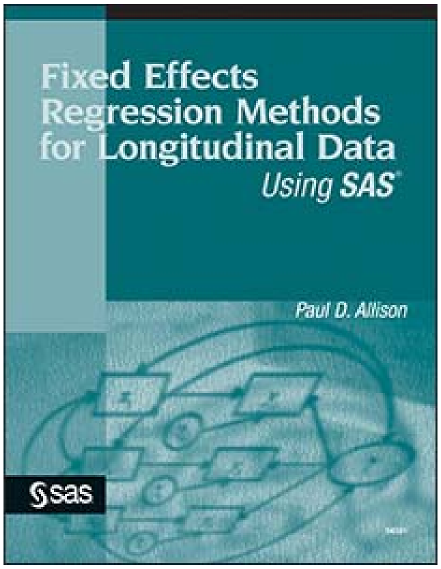
To control for unobservables, we can used *fixed effects* methods

- 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** or **xtlogit** in Stata
- 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


FIXED EFFECTS REGRESSION MODELS
Paul D. Allison
Series: Quantitative Applications in the Social Sciences
160
SAGE

2009

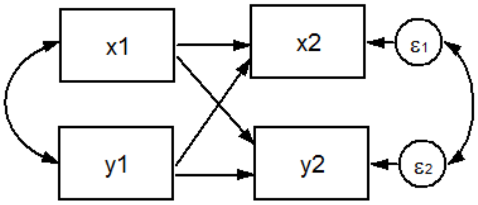


2005

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:

$$y_{it} = \mu_t + \overbrace{\beta_1 x_{i,t-1}}^{\text{Cross-lagged Effect}} + \beta_2 y_{i,t-1} + \delta_1 w_{it} + \gamma_1 z_i + \overbrace{\alpha_i}^{\text{Fixed Effect}} + \varepsilon_{it}$$

$$x_{it} = \tau_t + \beta_3 x_{i,t-1} + \underbrace{\beta_4 y_{i,t-1}}_{\text{Cross-lagged Effect}} + \delta_2 w_{it} + \gamma_2 z_i + \underbrace{\eta_i}_{\text{Fixed Effect}} + \upsilon_{it}$$

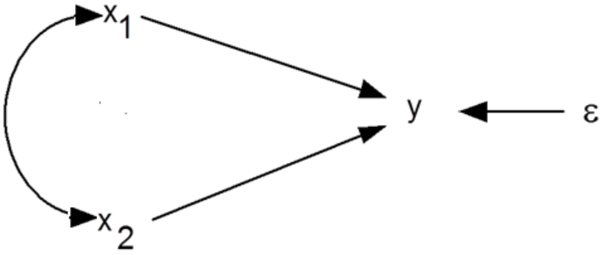
To get there, we'll

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

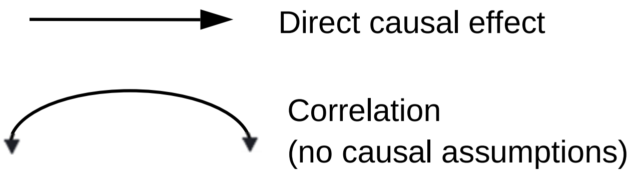
Path Analysis of Observed Variables

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.

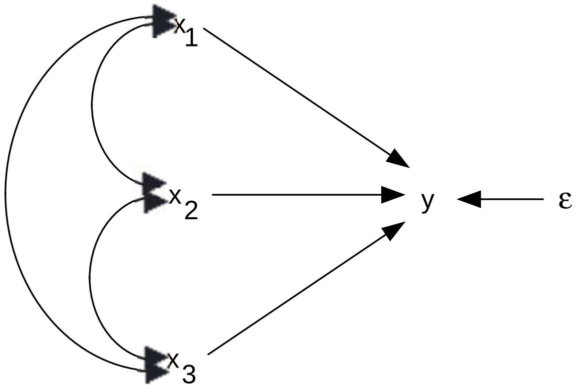
<u>Endogenous variables:</u> Variables caused by other variables in the system. These variables have straight arrows leading into them.

<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



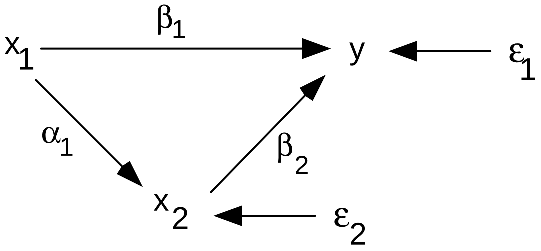
The fact that there are no curved arrows between ε and the x's implies that $\rho_{1\varepsilon} = 0$, $\rho_{2\varepsilon} = 0$, and $\rho_{3\varepsilon} = 0$. We make this assumption in the usual linear regression model.

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



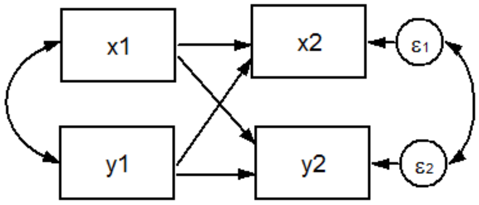
Note: The diagram goes further than the equations by asserting that

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

Cross-Lagged Linear Models

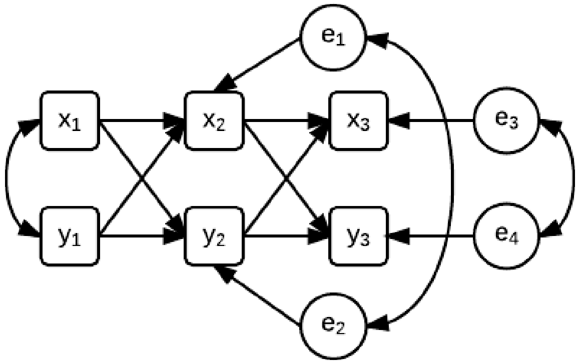
$y_2 = b_0 + b_1 y_1 + b_2 x_1 + \varepsilon_2$

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



- 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).

<u>Time-varying variables (measured at each of the three time points)</u>:

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.

<u>Time-invariant variables</u>:

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., ANTI90, ANTI92 and ANTI94.

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 the **sem** command.

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

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

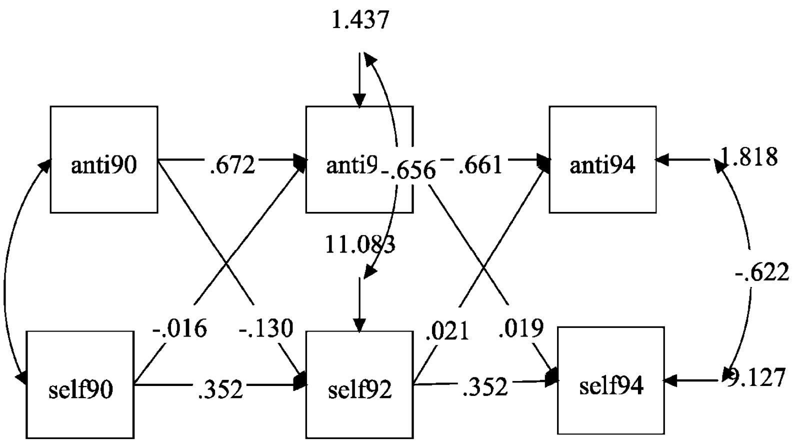
```
|                              OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  anti94 <-  |
      anti92 |   .6606303   .0365617    18.07   0.000     .5889707    .7322899
      self92 |   .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.

Path Diagram



Estimation & Assumptions

By default, **sem** does 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 test has approximately a chi-square distribution. The df is equal to the number of overidentifying restrictions.
- 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.
- Note that, although this statistic is properly regarded as a test of the model, 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.

Other Measures of Fit

```
. estat gof, stats(all)
----------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+------------------------------------------------
Likelihood ratio     |
          chi2_ms(4) |     48.074   model vs. saturated
            p > chi2 |      0.000
         chi2_bs(14) |    800.942   baseline vs. saturated
            p > chi2 |      0.000
---------------------+------------------------------------------------
Population error     |
               RMSEA |      0.138   Root mean squared error of
                     |                                 approximation
 90% CI, lower bound |      0.104
         upper bound |      0.174
              pclose |      0.000   Probability RMSEA <= 0.05
```

We want the RMSEA to be < .05. Definitely don't want it to be > .10

Other Measures of Fit

```
---------------------+------------------------------------------------
Information criteria |
                 AIC |  14925.739   Akaike's information criterion
                 BIC |  15004.304   Bayesian information criterion
---------------------+------------------------------------------------
Baseline comparison  |
                 CFI |      0.944   Comparative fit index
                 TLI |      0.804   Tucker-Lewis index
---------------------+------------------------------------------------
Size of residuals    |
                SRMR |      0.034   Standardized root mean squared residual
                  CD |      0.468   Coefficient of determination
----------------------------------------------------------------------
```

We want the CFI and TLI to be close to 1, definitely not below .90.

Global Goodness of Fit Measures

We want a single number that measures the similarity of $\hat{\mathbf{\Sigma}}$ and $\mathbf{S}$, the predicted covariance matrix (based on the model) and the observed covariance matrix.

As a general approach to model evaluation, LR chi-square may be too sensitive to sample size. Many alternative statistics have been proposed. Here are some that are reported by Stata.

Tucker Lewis Index (TLI)

Also known as Bentler & Bonnet's NonNormed Fit Index

Let χ_1^2 be the chi-square for the fitted model and let χ_0^2 be the chi-square for some baseline model, usually the "independence" model which says all the observed covariances are really 0. (Mplus only considers covariances among endogenous variables and between endogenous and exogenous variables).

$$TLI = \frac{\dfrac{\chi_0^2}{df_0} - \dfrac{\chi_1^2}{df_1}}{\dfrac{\chi_0^2}{df_0} - 1}$$

Adjusts for relative complexity of the two models. Can sometimes be greater than 1, a possible indication of "overfitting".

Other Global Measures

Comparative fit index

$$CFI = \frac{\left(\chi_0^2 - df_0\right) - \left(\chi_1^2 - df_1\right)}{\chi_0^2 - df_0}$$

As with the TLI, models pay a penalty for more parameters. The formula can be greater than 1 or less than 0, in which case the CFI is simply set to 1 or 0.

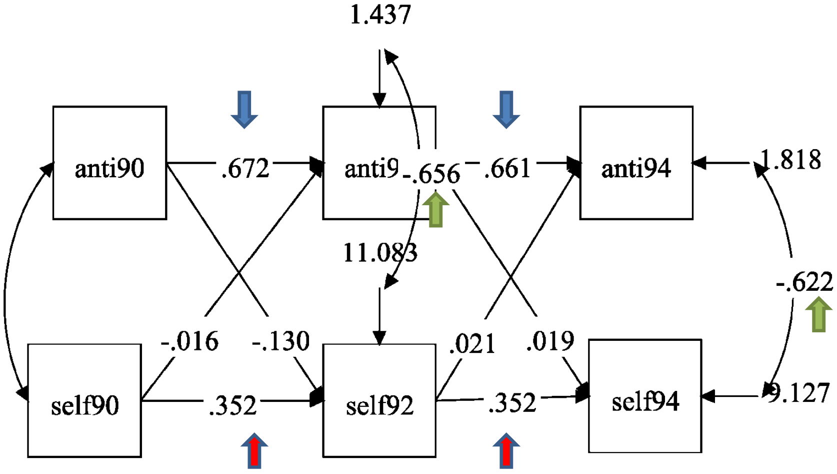
Root mean squared error of approximation:

$$RMSEA = \sqrt{\frac{\dfrac{\chi_1^2}{df_1} - 1}{N - 1}}$$

Good models have an RMSEA of .05 or less. Models whose RMSEA is .10 or more have poor fit. One nice thing about this statistic is that you can get a confidence interval.

Equality Constraints

If *T* > 2, consider constraining effects to be equal across time



Why? Get smaller standard errors -> narrower confidence intervals and more powerful hypothesis tests. Also easier to interpret.

Stata Program with Constraints

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

Compare this model with the last one:

Chi-square: 52.4 – 48.0 = 4.4

DF: 9 – 4 = 5

p = .49

So imposing the constraints did not significantly worsen the fit.