Structural Equation Models Using Stata

Rose Medeiros

Stata Webinar
March 18, 2020

Goals

- Learn a bit about structural equation modeling (SEM) and generalized structural equation modeling (GSEM)
- Learn about Stata's facilities for fitting SEM and GSEM

Stata's Tenets

- "Stata" rhymes with "data"
- Type a little, get a little
  - Click a little, get a little works fine, also
- Simple reproducibility
- Easy and complete extensibility
- Easy sharing

Notes and Slides

- Every slide in the presentations is in the notes
- The output from commands is only in the notes

Typography

- For commands, there are various fonts which can be used
  - `Items which must be typed as shown` will be in a `monospaced font`
  - *Items for which a substitution is needed* will be in *italics*
  - [*Optional items*] will be [*in square brackets*], though the brackets do not get typed
- Example Stata commands will often be preceded by a `.`
  - The `.` is a prompt and does not get typed—it is for distinguishing input from output in the notes

Descriptions of Linear SEM

- SEM is a class of statistical techniques that allows us to test hypotheses about relationships among variables
- SEM may also be referred to as Analysis of Covariance Structures
  - SEM fits models using the observed covariances and, possibly, means
- SEM encompasses other statistical methods such as correlation, linear regression, and factor analysis

Descriptions of Linear SEM (Continued)

- SEM is a multivariate technique that allows us to estimate a system of equations
  - Variables in these equations may be measured with error
  - There may be variables in the model that cannot be measured directly

SEM in Stata

- The `sem` command is used to fit standard linear SEM models
- The `gsem` command is used to fit generalizations of linear SEM
  - Generalized linear models including continuous, binary, ordinal, nominal, count, and survival outcomes
  - Multilevel models including random intercepts, random slopes, and crossed effects
- Both types of models can be specified using either commands or path diagrams via the SEM Builder

Types of Variables

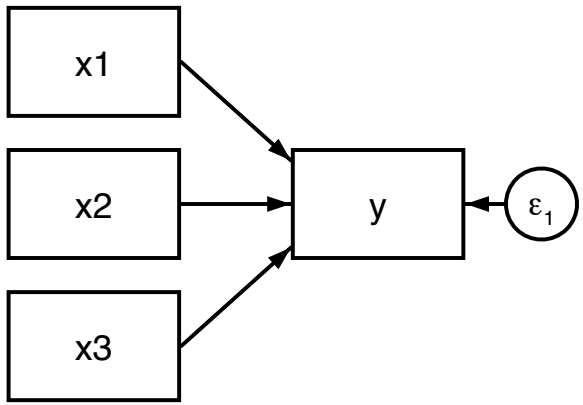
- Observed variables (manifest variables)
  - We have observed values of these variables in our dataset
  - Or we have variances, covariances, and possibly means based on observed values
- Latent variables (unobserved variables, factors)
  - May represent the following
    - Hypothetical constructs
    - Variables that cannot be directly measured
    - True values of variables measured with error
    - Unobserved heterogeneity
    - Errors or disturbances
  - Latent variables are measured by observed variables known as measurements or indicators

Types of Variables (Continued)

- Endogenous variables
  - Also known as y, dependent, or response variables
  - Predicted by one or more other variables
  - May predict other endogenous variables
- Exogenous variables
  - Also known as x, independent, or explanatory variables
  - Variables that are not predicted by any other variables in the model
  - May be correlated with other variables in the model

Path Diagrams

- Observed variables are represented by rectangles
- Latent variables are represented by ovals or circles
- Paths are represented by arrows
- Covariances are represented by curved lines with arrows at each end
- The path diagram below corresponds to a linear regression of y on x1, x2, and x3



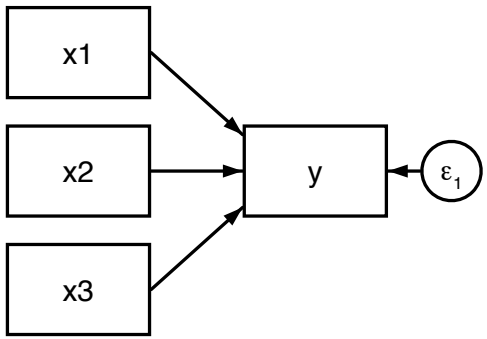
Data

- `sem` allows two types of data
- Datasets with individual observations
- Datasets made up of summary statistics, specifically covariance or correlation matrices; and possibly means
  - See `help ssd` for more information about working with summary statistics data

Basic Syntax

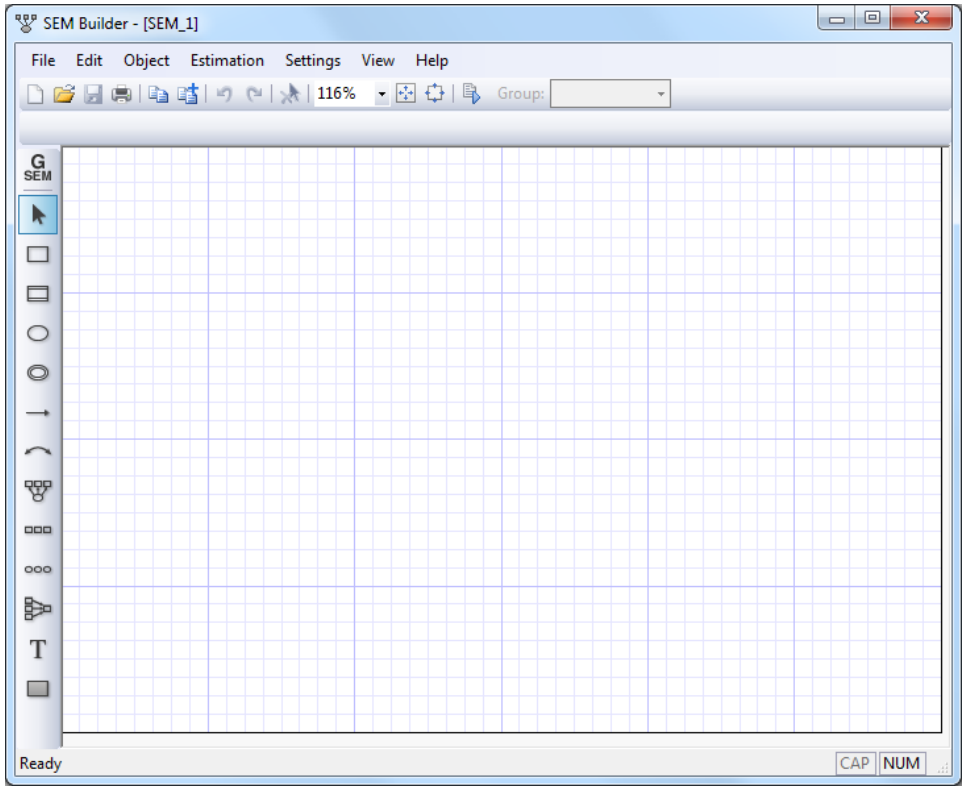
- `sem` *paths* [`if`] [`in`] [`weight`] [`, options`]
- The basic rules are
  - All paths are placed inside parentheses
  - Arrows point towards dependent variables
- Beyond that the *paths* specifications are flexible
- The three commands below all fit the path model shown here

```
. sem (y <- x1 x2 x3)
. sem (x1 x2 x3 -> y)
. sem (x1 -> y) (x2 -> y) (x3 -> y)
```



The SEM Builder

- To open the SEM Builder, type `sembuilder` or click on **Statistics > SEM (structural equation modeling) > Model building and estimation**



- The tools along the left-hand side allow us to draw the path

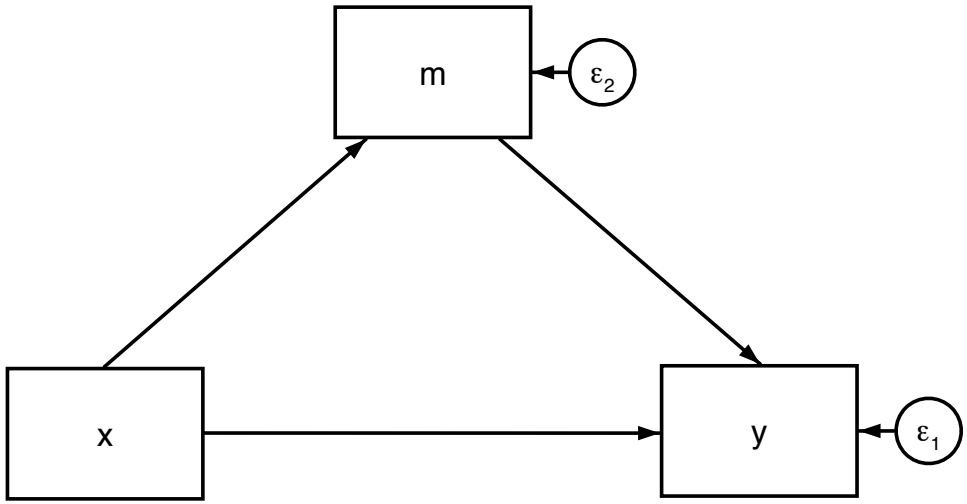
Path Analysis

- We will begin by looking at some examples of path analysis
- Path models include only observed variables and their error terms
- These models can be simple or they may include many observed variables with intricate relationships

Mediation Models

- In a mediation model, a variable `x` is hypothesized to predict `y` in two ways
  - Directly
  - Indirectly because `x` predicts a third variable `m` which in turn predicts `y`
- Using `sem` we can fit all of the equations in the mediation model simultaneously

A Simple Mediation Model



- The command for the above model is

```
. sem (m <- x) (y <- x m)
```

Job Satisfaction Data

- Fogarty et al. (1999) fit a variety of models that examine relationships among positive and negative affectivity, stress, coping, strain, and job satisfaction
  - We'll fit a much simpler model
- The data for this example are stored as summary statistics in `jobsat.dta`

```
. use jobsat
```

- We can learn about the variables

```
. ssd describe
```

- We can also look at their summary statistics

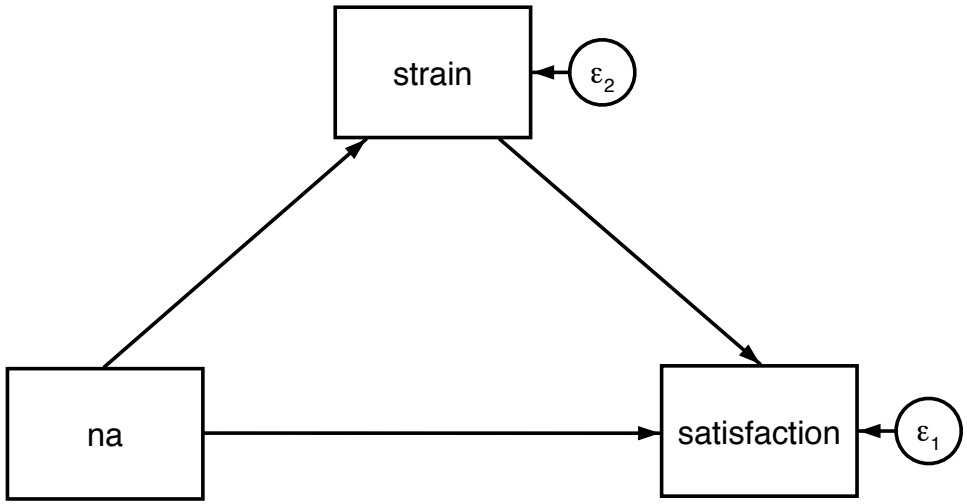
```
. ssd list
```

Fitting a Mediation Model

- As an example, we can fit a model where strain mediates the effect of negative affectivity on job satisfaction

```
. sem (satisfaction <- strain na) (strain <- na)
```

- In path diagram form the model is



Direct, Indirect, and Total Effects

- Direct effects are the coefficient estimates we see in the output
  - The direct effect of `na` on `satisfaction` is .2
- We can calculate the indirect effect of negative affectivity on job satisfaction by multiplying the appropriate coefficients
  - The path coefficient from `na` to `strain` is roughly 1.62
  - The path coefficient from `strain` to `satisfaction` is roughly $-.322$
  - $1.62 \times -.322 = -.522$ so the indirect effect is roughly $-.52$
- The total effect is the sum of the direct and indirect effects
  - The total effect of negative affectivity on job satisfaction is $.2 + (-.52) = -.32$

estat teffects

- We can obtain direct, indirect, and total effects and their standard errors using `estat teffects`

```
. estat teffects
```

- What do we see?
  - The direct effect of `na` on `satisfaction` is not significantly different from 0
  - The direct effect of `strain` on `satisfaction` is significant different from 0
  - The indirect effect of `na` on `satisfaction` is also significantly different from 0
  - The direct effect of `na` on `satisfaction` can be said to be fully mediated by `strain`

Standardized coefficients

- We can replay the model with standardized coefficients using

  ```
  . sem, standardized
  ```

- We expect that a one standard deviation increase in strain would produce a .58 standard deviation decrease in satisfaction, holding negative affectivity constant
- We can also obtain standardized indirect and total effects

  ```
  . estat teffects, standardized nodirect
  ```

  - The standardized indirect effects are the products of standardized coefficients

Equation Level Goodness-of-fit

- We can also obtain variance decomposition and R-squared for each of the endogenous variables in the model

```
. estat eqgof
```

- About 27% of variation in `satisfaction` is explained by the model
- To obtain a Wald test for the null hypothesis that all coefficients in an equation are 0 we can use

```
. estat eqtest
```

Nonrecursive Path Models

- The examples so far have been of recursive models
- Models that contain feedback loops or correlated error terms are said to be nonrecursive
- `sem` can be used to fit nonrecursive models
- Checking whether a nonrecursive model is identified can be simple or complex, depending on the model
- Stata provides tools for evaluating identification of nonrecursive models, for more information see `help estat stable`

Models with Latent Variables

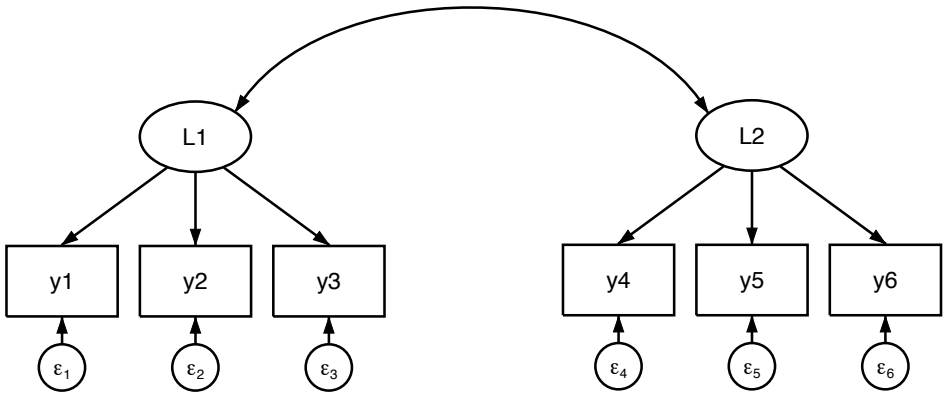
- Models with latent variables can estimate and accommodate measurement error
- Some variables cannot be directly measured without error
  - But we may be able to collect multiple measurements each of which contains some error
- Confirmatory factor analysis (CFA) allows us to evaluate how well the items we collect measure the corresponding concept
- CFA models are also called measurement models
- Examples: Personality features, depression, attitudes

Confirmatory Factor Analysis

- In a confirmatory factor analysis (CFA) model, one or more latent variables is measured by a series of observed variables
  - The latent variables may be correlated, but no structural paths are specified
- Each latent variable is associated with a set of observed variables
- These models are confirmatory in the sense that we specify them based on prior knowledge or theory about
  - What the latent variables represent
  - Which observed variables are associated with each latent variable
- This is unlike exploratory factor analysis where all observed variables are allowed to measure each of the latent variables

Confirmatory Factor Analysis (continued)

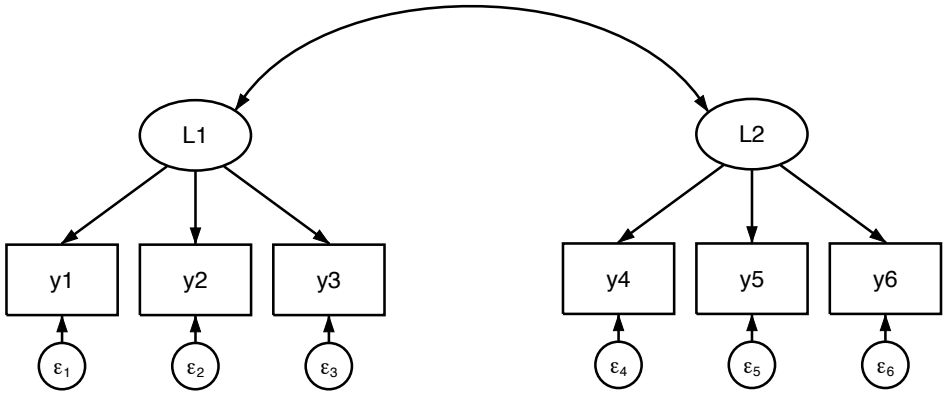
- At least 3 measurement variables are required to identify a model that contains only a single latent variable
- Sometimes also called a measurement model
- Here is an example of a path diagram for a CFA model



Syntax for a CFA Model

- By default, variables with names beginning with a capital letter are assumed to be latent variables
- Three equivalent methods of fitting the CFA model shown below are

```
. sem (L1 -> y1 y2 y3) (L2 -> y4 y5 y6)
. sem (y1 y2 y3 <- L1) (y4 y5 y6 <- L2)
. sem (L1 -> y1) (L1 -> y2) (L1 -> y3) ///
.         (L2 -> y4) (L2 -> y5) (L2 -> y6)
```



Education Data

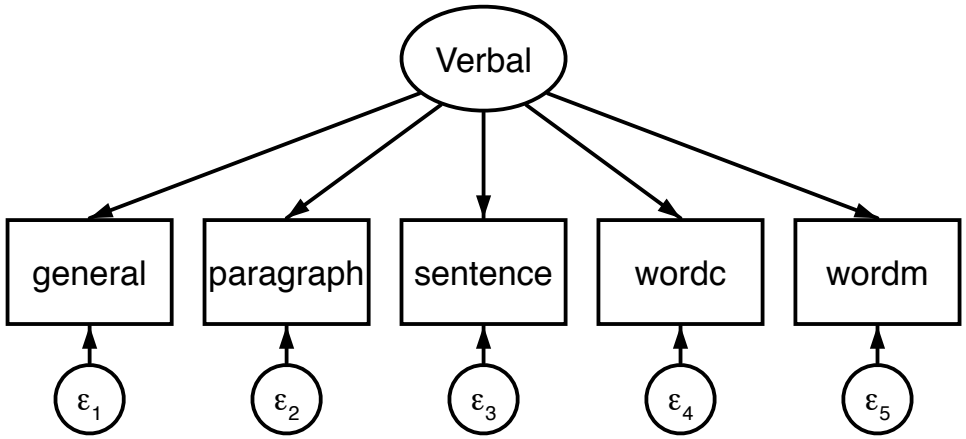
- Holzinger and Swineford (1939) measured the abilities of students across a variety of areas
- Five of the observed variables measure verbal skills
- Let's open the data and take a look at these variables

```
. use hsdata
. codebook general paragraph sentence wordc wordm
```

- We'll use the observed variables `general`, `paragraph`, `sentence`, `wordc`, and `wordm` as indicators for the latent variable representing verbal abilities

Fitting a Single Factor CFA

- The latent variable, Verbal, is assumed to cause the observed variables



- The `sem` command to fit the model is

```
. sem (Verbal -> general paragraph sentence wordc wordm)
```

Fitting a Single Factor CFA (Continued)

- For each observed variable we obtain an intercept and path coefficient
  - The path from the latent variable to the first observed variable is constrained to 1 for identification
- The error variances for our indicators represent the portion of the indicator's variance that is not explained by the latent variable `Verbal`
- The overall model χ^2 tests the null hypothesis that the covariance matrix implied by our model is equal to the observed covariance matrix in the population
  - If the model includes means, the respective mean vectors are included in this test

Examining Model Fit

- We can examine the fit of this model using

  ```
  . estat gof, stats(all)
  ```

- The first χ^2 test is the same test reported in the output from `sem`
- The second χ^2 compares the saturated model with a baseline model that includes
  - The means and variances of all observed variables, and
  - The covariances of all observed exogenous variables
  - Different authors define the baseline model differently

More Model Fit

- A variety of measures of fit have been proposed for SEM
  - For many of these measures a variety of standards for what constitutes good or acceptable fit have also been proposed
- `sem` provides the following
  - RMSEA or root mean square error of approximation
    - The p-value labeled pclose corresponds to a test of RMSEA$< .05$
  - AIC and BIC
  - The comparative fit index (CFI) and Tucker-Lewis index (TLI)
  - Standardized root mean squared residual (SRMR)
  - Coefficient of determination (CD)

Standardized Coefficients

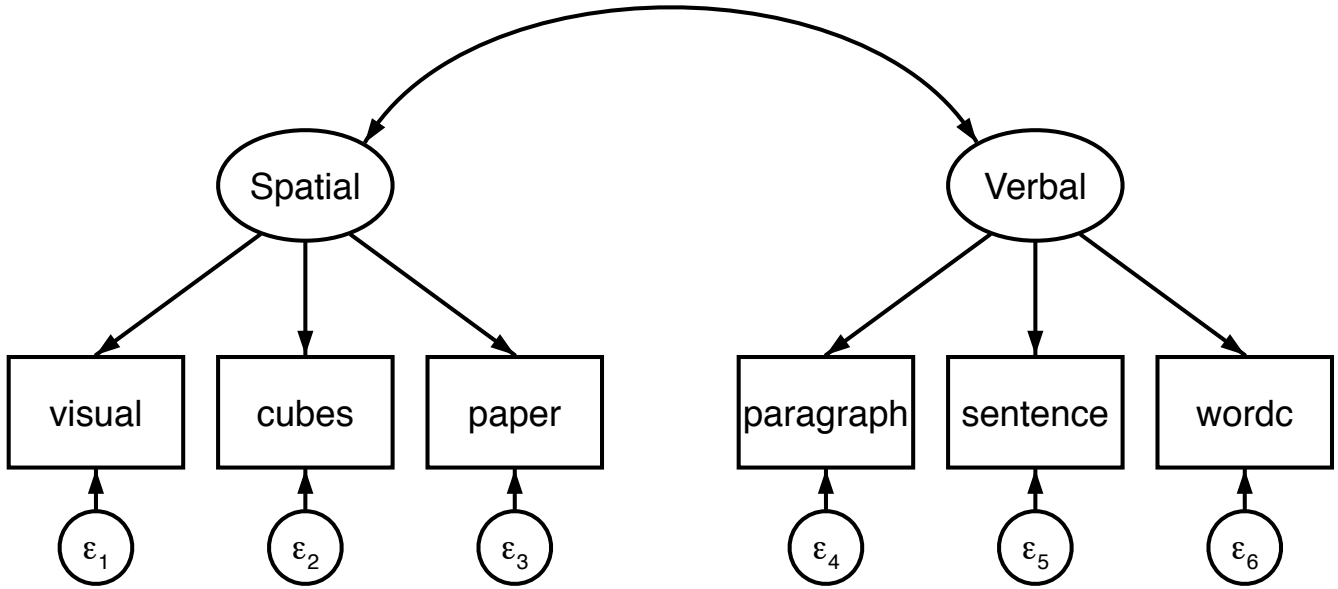
- We can replay the model to obtain standardized coefficients

```
. sem, standardized
```

- The standardized loading is the correlation between the latent variable and the observed variable when each indicator measures only a single latent variable
- The standardized error variances are the proportion of variation not explained by the latent variable

Two Factor CFA

- Now we'll add a second latent variable called `Spatial` which represents students' spacial abilities, using the indicators `visual`, `cube`, and `paper`



- Read nothing into the fact that we have reduced the number of items for the variable `Verbal` from five to three

Fitting a Two Factor CFA

- We can fit this model using

```
. sem (Spatial -> visual cubes paper) ///
      (Verbal -> paragraph sentence wordc)
```

- We can examine model fit using

```
. estat gof, stats(all)
```

Modification Indices

- MIs are used to check for paths and covariances that could be added to the model to improve model fit
  - Over-fitting is a serious danger here
- Approximate change in the χ^2 statistic if the parameter is added to the model
- To obtain modification indices for our model we can type

```
. estat mindices
```

- The largest MIs are associated with
  - Adding a path from `Spatial` to `sentence`
  - Adding a covariance between the error terms for `paragraph` and `wordc` (word classification)

Refitting our Model

- We can use the `var()` option to add the suggested covariance between error terms
- In this case we use `var(e.paragraph*e.wordc)`

```
. sem (Spatial -> visual cubes paper) ///
      (Verbal -> paragraph sentence wordc), ///
      var(e.paragraph*e.wordc)
```

- In the SEM Builder, you can add a curved, double-headed arrow between `paragraph` and `wordc`
- After fitting the model we could check model fit to see if our model fits better. We could also check the MIs again.

Full Structural Equation Models

- Combine path analysis and confirmatory factor analysis
- One or more latent variables are included in the model with corresponding observed indicators
- Structural relationships may exist among latent and/or observed variables

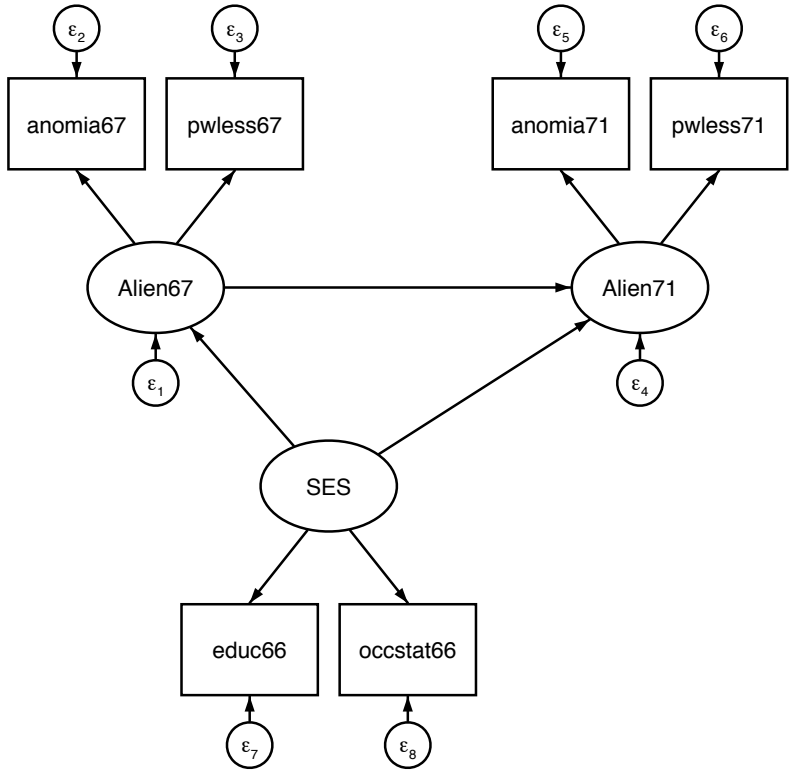
Data on Alienation

- The data for this set of examples come from Wheaton et al. (1977)

```
. use wheaton
. ssd describe
```

- The model includes three latent variables with structural paths between them
- Alienation in 1971 is predicted by alienation in 1967 and socioeconomic status in 1966
- In each year, alienation is measured by observed variables measuring feelings of anomia and powerlessness
- Socioeconomic status is measured by education level and occupational status

The Alientation Model



Fitting a SEM Model

- Fit the model

```
. sem (Alien67 -> anomia67 pwless67) ///
      (Alien71 -> anomia71 pwless71) ///
      (SES -> educ66 occstat66) ///
      (Alien67 <- SES) ///
      (Alien71 <- Alien67 SES)
```

Revising the Model

- On substantive grounds we could argue that covariances should be added between the errors for
  - Powerlessness in 1967 and 1971
  - Anomia in 1967 and 1971
- One method of evaluating whether adding these covariances to the model improves model fit is to perform a likelihood-ratio test
  - In the SEM literature this is often called the χ^2 difference test
- We can use the `lrtest` command to perform this test
- First we will need to fit both models

The Likelihood-ratio Test

- We'll start by storing the estimates from the model we have already run

```
. estimates store nocov
```

- Now we can run the model with the covariances added

```
. sem (Alien67 -> anomia67 pwless67) ///
      (Alien71 -> anomia71 pwless71) ///
      (SES -> educ66 occstat66) ///
      (Alien67 <- SES) ///
      (Alien71 <- Alien67 SES), ///
          cov(e.anomia67*e.anomia71 ///
             e.pwless67*e.pwless71)
```

- First we store the estimates

```
. estimates store withcov
```

The Likliehood-ratio Test (Continued)

- Then we can run the likelihood ratio test

```
. lrtest nocov withcov
```

- The likelihood-ratio test indicates significantly better fit with the two covariances
- We could also have run only the model with the covariances and used the `test` command to perform a Wald test for joint significance of the covariances

Comparing Groups

- Multiple group SEM allows for estimating parameters of a model separately for across groups
  - All parameters may be estimated separately, or
  - Some or all parameters can be constrained to equality across groups
- This allows us to examine whether parameters vary across groups
- We can use the `group()` option of `sem` to fit a model for two or more groups
- The `ginvariant()` option allows you to specify what parameters should be constrained across groups
  - By default measurement coefficients and measurement intercepts are constrained across groups

Multiple Group CFA

- To demonstrate we will return to the two-factor CFA model we fit earlier

```
. use hsdata, clear
```

- The students in this dataset come from two different schools

```
. tab school
```

A Model with No Cross-group Constraints

- We will begin by estimating all parameters separately for each group to do this we will
  - Specify `ginvariant(none)`
  - Set the means of the latent variables to 0 using the `mean()` option
- Our command is

```
. sem (Spatial -> visual cubes paper) ///
      (Verbal -> paragraph sentence wordc), ///
     mean(Spatial@0 Verbal@0) ///
     group(school) ginvariant(none)
```

Testing for Group Differences

- We could use a likelihood-ratio test to see whether the parameters vary across schools
- An alternative is to use `estat ginvariant` instead
  - Wald tests are used to test whether *unconstrained* coefficients are significantly different
  - Score tests are used to test whether *relaxing constraints* would improve model fit
  - In both cases the null hypothesis is that the constraint does not harm model fit
- Let's give it a try

```
. estat ginvariant
```

- We fail to reject the null hypothesis that the measurement coefficients are the same across groups

Constraining Coefficients

- We could refit the model constraining measurement coefficients to equality across gourps using the option `ginvariant(mcoef)`
- Then we could use `estat ginvariant` to test for equality of the intercepts across groups
- A typical ordering of tests is
  - Measurement coefficients
  - Measurement intercepts
  - Measurement error variances

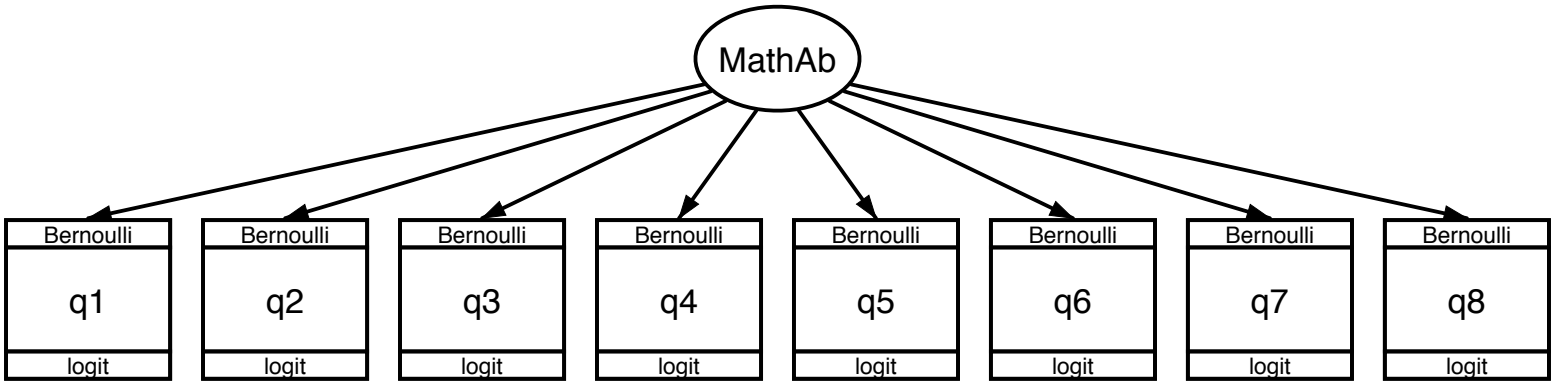
Generalized Structure Equation Models

- `gsem` allows us to extend the types of models we can fit
- Models for binary, ordered, nominal, count, survival time, interval, and censored responses.
- Multilevel models, including models with random intercepts and slopes, for nested or crossed data
- Latent variables can be included at any level of the model

CFA with Binary Indicators

- Many of the models we fit above can be extended to include generalized response variables
- In this example we will fit a confirmatory factor analysis model using binary indicators
- The dataset contains fictional data on students' math scores and attitudes towards math
- The binary indicators are 8 questions from a math test
- The latent variable is math ability

Path Model for a Generalized CFA



Fitting a Generalized CFA

- Let's begin by opening the dataset and looking at the items `q1-q8`

```
. use math
. codebook q1-q8
```

- There are only a few changes to the command
  - We will use the `gsem` command to fit this model
  - Specify the `logit` option to fit a logit model to our binary response variables `q1-q8`
  - We'll use the `nodvheader`, otherwise `gsem` will list the family and link function for each dependent variable

```
. gsem (MathAb -> q1-q8, logit), nodvheader
```

- If we include certain constraints on this model, it can be interpreted as an item response theory (IRT) model
  - See `help irt` for information on Stata's `irt` commands

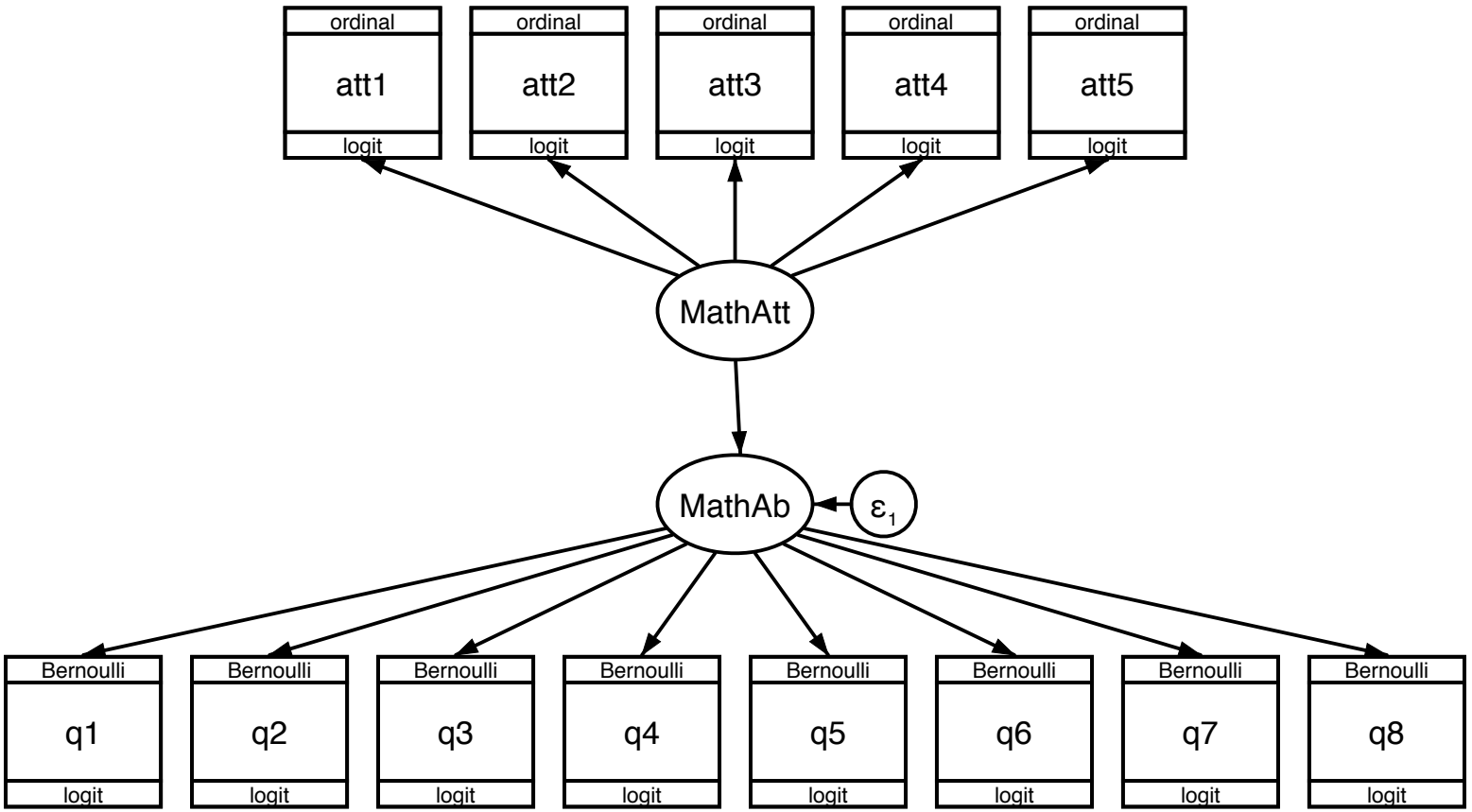
A Generalized Structural Equation Model

- The dataset also includes information on student's attitudes towards math, we may want to see if these predict math ability
- Let's look more closely at these items

```
. codebook att*
```

- The math attitude items appear to be Likert-type items

Path Diagram for the GSEM



Fitting a GSEM

- We will use the `ologit` option to model the responses `att1-att5` using an ordered logistic model

```
. gsem (MathAb -> q1-q8, logit) ///
       (MathAtt -> att1-att5, ologit) ///
       (MathAtt -> MathAb), nodvheader
```

Multilevel SEM

- Many of the types of structural equation models that we have discussed can be extended to multilevel models using `gsem`
- Because `gsem` can be used to include random effects and model generalized responses a large number of models can be fit
  - Including a multilevel multinomial logit model that cannot be fit elsewhere

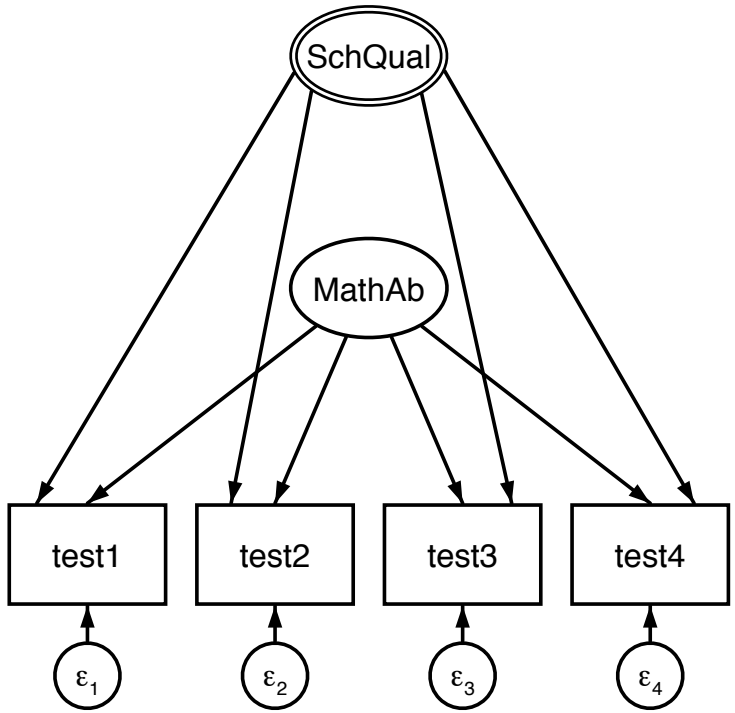
Multilevel CFA

- We will look at an example of a multilevel CFA
- We will continue using the the same dataset
- The students are clustered within schools
- This time we will measure the latent variable `MathAb` using test scores, let's take a look

```
. codebook school test*
```

Path Diagram for a Multilevel CFA

- Random effects are denoted as ovals with double rings
- Graphically the model is



Fitting a Multilevel CFA

- Here the test items are predicted by `MathAb` and the random intercept denoted `SchQual`
- The square brackets around school indicate that `SchQual` is constant within school and varies across schools
- Run the model

```
. gsem (MathAb SchQual[school] -> test1 test2 test3 test4)
```

Conclusion

- We have learned a bit about structural equation models and generalized structural equation models
- We have seen how to use `sem` to fit linear SEM models
- We have also seen how `gsem` can be used to fit more general models
- We have also touched on the flexibility of `gsem`