

# Introduction to Structural Equation Modeling Using Stata

Chuck Huber  
StataCorp

California Association for Institutional Research  
November 19, 2014

# Outline

- Introduction to Stata
- What is structural equation modeling?
- Structural equation modeling in Stata
- Continuous outcome models using **sem**
- Multilevel generalized models using **gsem**
- Demonstrations and Questions

# Introduction to Stata

- The Stata interface
- The menus and dialog boxes
- Stata command syntax
- The data editor
- The do-file editor

# The Stata Interface

The screenshot displays the Stata software interface. The main window shows the command window with the following content:

```

. describe id university college private fygpa ret_yr1

```

variable name	storage type	display format	value label	variable label
<b>id</b>	int	%9.0g		<b>Identification Number</b>
<b>university</b>	byte	%9.0g		<b>University ID</b>
<b>college</b>	byte	%11.0g	college	<b>Primary college of major</b>
<b>private</b>	byte	%9.0g	private	<b>Private or public university?</b>
<b>fygpa</b>	double	%4.2f		<b>First-year GPA</b>
<b>ret_yr1</b>	byte	%8.0g	YesNo	<b>* First-year retention</b>

The right-hand side of the interface features two panels:

- Variables:** A list of variables with their labels. The selected variable is `ret_yr1` (First-year retention).
- Properties:** A detailed view of the selected variable's properties:
 

Properties	
Name	ret_yr1
Label	First-year retention
Type	byte
Format	%8.0g
Value Label	YesNo
Notes	
Data	
Filename	cair.dta
Label	Example data for th...
Notes	
Variables	26
Observations	12,958
Size	1.25M

The bottom status bar shows the current directory as `C:\NC120` and the current view as `CAP NUM OVR`.

# The Menus and Dialog Boxes

The screenshot displays the Stata software interface. The main window title is "Stata/MP - C:\Users\jch\Dropbox\Talks\2014\CAIR\examples\cair.dta - [Results]". The menu bar includes File, Edit, Data, Graphics, Statistics, User, Window, and Help. The Statistics menu is open, showing a list of statistical procedures. The "Linear models and related" option is selected, which has opened a sub-menu. In this sub-menu, "Linear regression" is selected. To the right, the "regress - Linear regression" dialog box is open. It has tabs for Model, by/f/in, Weights, SE/Robust, and Reporting. The "Model" tab is active. The "Dependent variable:" field is empty, and the "Independent variables:" field is also empty. Under the "Treatment of constant" section, there are three checkboxes: "Suppress constant term", "Has user-supplied constant", and "Total SS with constant (advanced)". At the bottom of the dialog box are "OK", "Cancel", and "Submit" buttons. In the background, the command window shows the following commands:

```
# Command _rc
1 use "C:\Users\...
2 ds
```

At the bottom right of the Stata window, a table displays the following information:

Type	int
Format	%9.0g
Value Label	
Notes	
Data	
Filename	cair.dta
Label	Example data for th
Notes	
Variables	26
Observations	12,958
Size	1.25M

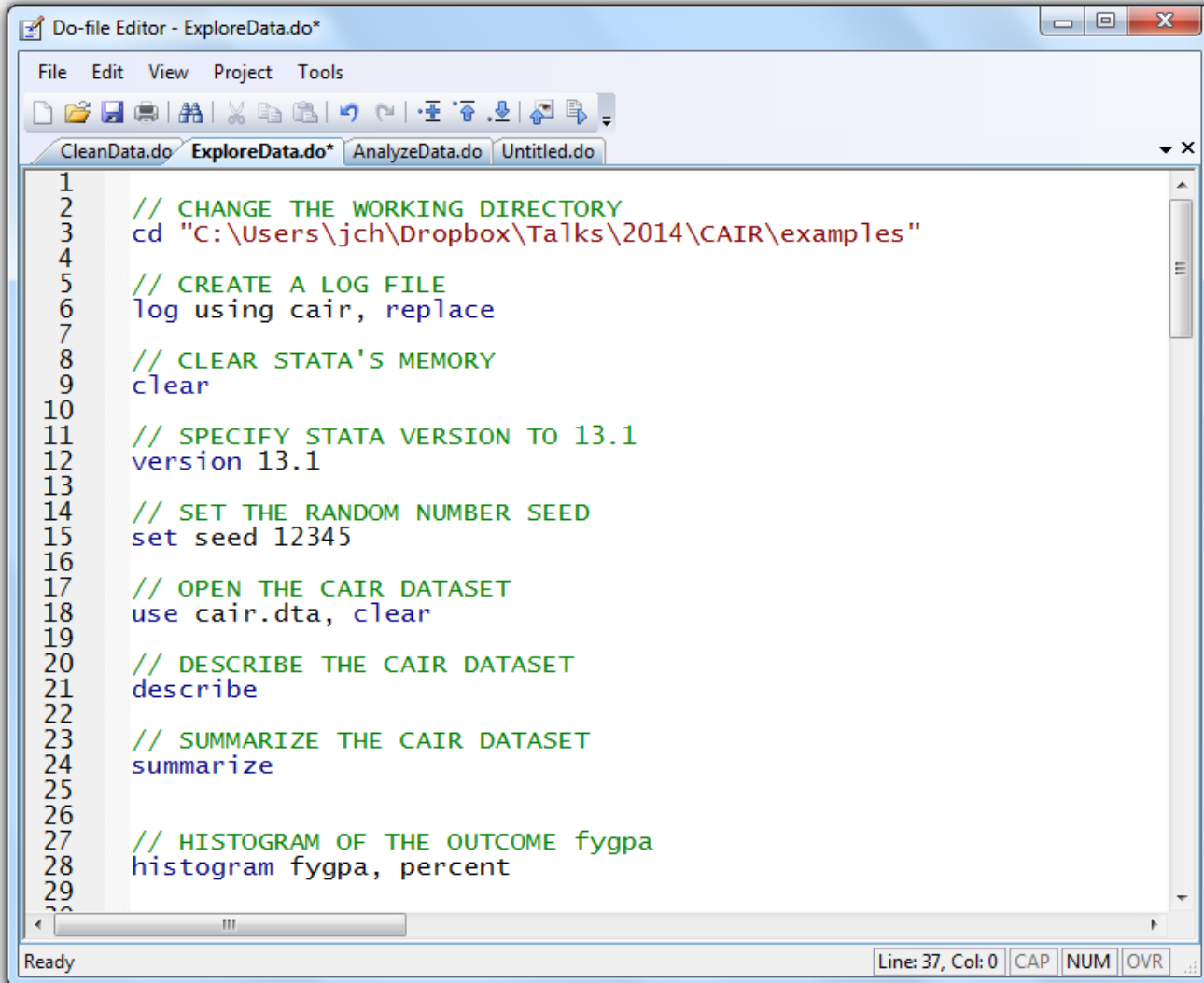
At the bottom left of the Stata window, the text "CAN120" is visible. At the bottom right, the status bar shows "CAP NUM OVR".

# The Data Editor

The screenshot shows the STATA Data Editor interface. The main window displays a data table with columns: id, university, college, private, fygpa, ret\_yr1, and instate. The 'id' variable is selected in the 'Variables' list on the right. The 'Properties' pane shows details for the 'id' variable, including its name, label, type, format, and value label. The 'Data' pane shows the filename 'cair.dta', label 'Example data for the Calif...', and other statistics like 26 variables and 12,958 observations.

	id	university	college	private	fygpa	ret_yr1	instate
1	1	20	Science	public	2.29	Yes	In-State
2	2	20	Science	public	2.65	Yes	In-State
3	3	14	Science	public	2.19	Yes	In-State
4	4	14	Education	public	2.85	Yes	In-State
5	5	12	Business	public	1.96	Yes	In-State
6	6	6	Science	private	3.05	Yes	In-State
7	7	10	Education	private	3.01	No	In-State
8	8	20	Science	public	2.30	Yes	In-State
9	9	12	Education	public	2.47	Yes	In-State
10	10	8	Education	public	3.16	Yes	In-State
11	11	20	Education	public	1.22	Yes	In-State
12	12	12	Education	public	1.87	Yes	In-State
13	13	2	Education	public	3.31	Yes	In-State
14	14	5	Education	private	2.61	Yes	In-State
15	15	19	Education	private	1.54	Yes	Out-of-State
16	16	4	Science	private	2.73	Yes	In-State
17	17	6	Science	private	1.48	Yes	In-State
18	18	8	LiberalArts	public	3.61	Yes	In-State
19	19	15	Engineering	public	2.77	Yes	In-State
20	20	18	Education	private	1.72	No	In-State
21	21	19	Science	private	2.54	Yes	Out-of-State

# The Do-File Editor



The screenshot shows the STATA Do-file Editor window titled "Do-file Editor - ExploreData.do\*". The window has a menu bar with "File", "Edit", "View", "Project", and "Tools". Below the menu bar is a toolbar with various icons for file operations. The main text area contains a script with the following content:

```
1
2 // CHANGE THE WORKING DIRECTORY
3 cd "C:\Users\jch\Dropbox\Talks\2014\CAIR\examples"
4
5 // CREATE A LOG FILE
6 log using cair, replace
7
8 // CLEAR STATA'S MEMORY
9 clear
10
11 // SPECIFY STATA VERSION TO 13.1
12 version 13.1
13
14 // SET THE RANDOM NUMBER SEED
15 set seed 12345
16
17 // OPEN THE CAIR DATASET
18 use cair.dta, clear
19
20 // DESCRIBE THE CAIR DATASET
21 describe
22
23 // SUMMARIZE THE CAIR DATASET
24 summarize
25
26
27 // HISTOGRAM OF THE OUTCOME fgpa
28 histogram fgpa, percent
29
30
```

The status bar at the bottom of the window shows "Ready" on the left and "Line: 37, Col: 0" on the right, along with buttons for "CAP", "NUM", and "OVR".

# Outline

- Introduction to Stata
- **What is structural equation modeling?**
- Structural equation modeling in Stata
- Continuous outcome models using **sem**
- Multilevel generalized models using **gsem**



# What is Structural Equation Modeling?

- Brief history
- Path diagrams
- Key concepts, jargon and assumptions
- Assessing model fit
- The process of SEM

# Brief History of SEM

- Factor analysis had its roots in psychology.
  - Charles Spearman (1904) is credited with developing the common factor model. He proposed that correlations between tests of mental abilities could be explained by a common factor representing ability.
  - In the 1930s, L. L. Thurston, who was also active in psychometrics, presented work on multiple factor models. He disagreed with the idea of a one general intelligence factor underlying all test scores. He also used an oblique rotation, allowing the factors to be correlated.
  - In 1956, T.W. Anderson and H. Rubin discussed testing in factor analysis, and Jöreskog (1969) introduced confirmatory factor analysis and estimation via maximum likelihood estimation, allowing for testing of hypothesis about the number of factors and how they relate to observed variables.

# Brief History of SEM

- Path analysis and systems of simultaneous equations developed in genetics, econometrics, and later sociology.
  - Sewall Wright, a geneticist, is credited with developing path analysis. His first paper using this method was published in 1918 where he looked at genetic causes related to bone sizes in rabbits. Rather than estimating only the correlation between variables, he created path diagrams to that showed presumed causal paths between variables. He compared what the correlations should be if the variables had the presumed relationships to the observed correlations to evaluate his assumptions.
  - In the 1930s, 1940s, and 1950s, many economists including Haavelmo (1943) and Koopmans (1945) worked with systems of simultaneous equations. Economists also introduced a variety of estimation methods and investigated identification issues.
  - In the 1960, sociologists including Blalock and Duncan applied path analysis to their research.

# Brief History of SEM

- In the early 1970s, these two methods merged.
  - Hauser and Goldberger (1971) worked on including unobservables into path models.
  - Jöreskog (1973) developed a general model for fitting systems of linear equations and for including latent variables. He also developed the methodology for fitting these models using maximum likelihood estimation and created the program LISREL.
  - Keesling (1972) and Wiley (1973) also worked with the general framework combining the two methods.
- Much work has been done since then in to extend these models, to evaluate identification, to test model fit, and more.

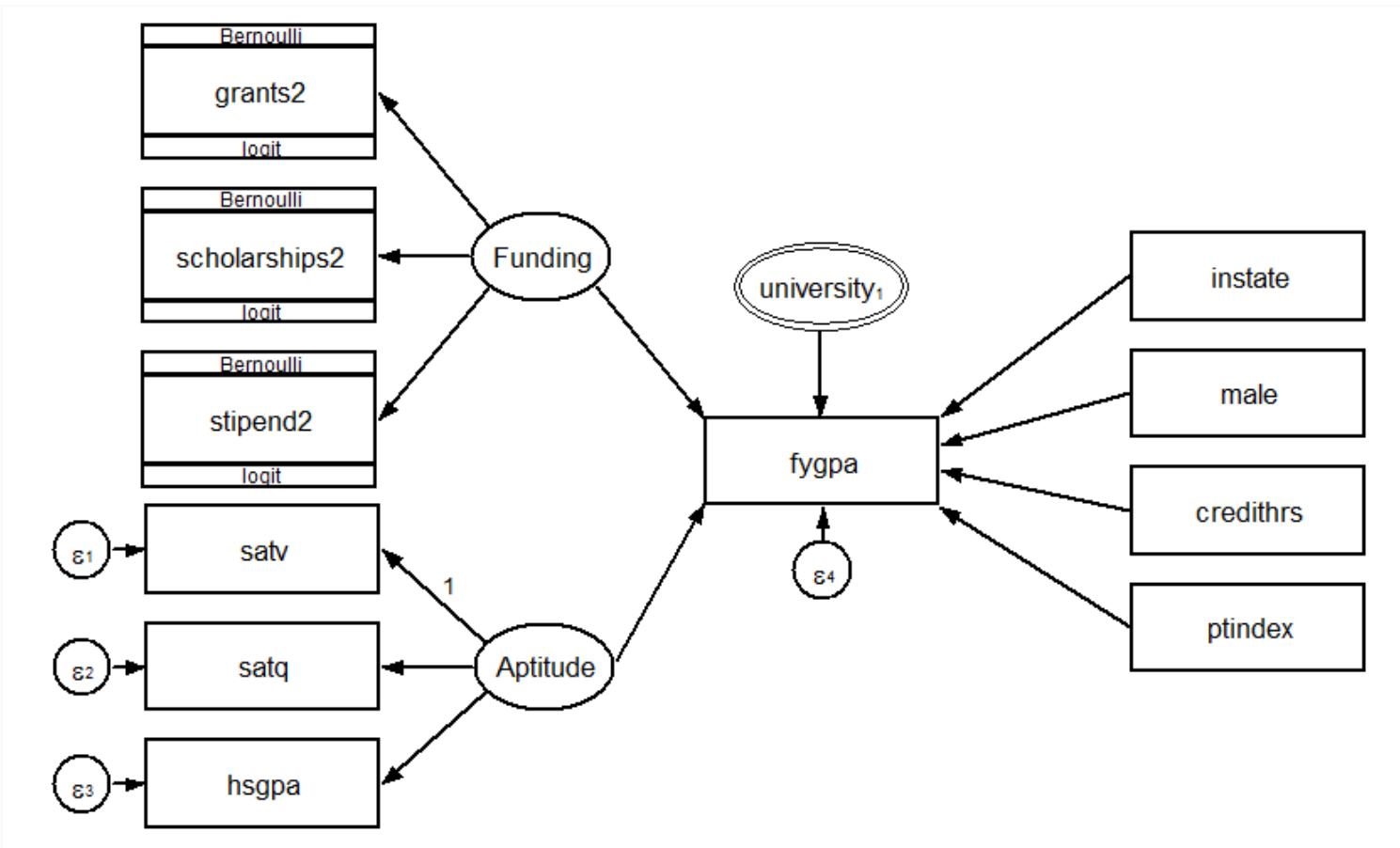
# What is Structural Equation Modeling?

- Structural equation modeling encompasses a broad array of models from linear regression to measurement models to simultaneous equations.
- Structural equation modeling is not just an estimation method for a particular model.
- Structural equation modeling is a way of thinking, a way of writing, and a way of estimating.

# What is Structural Equation Modeling?

- 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.
- 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.

# Structural Equation Models are often drawn as Path Diagrams:



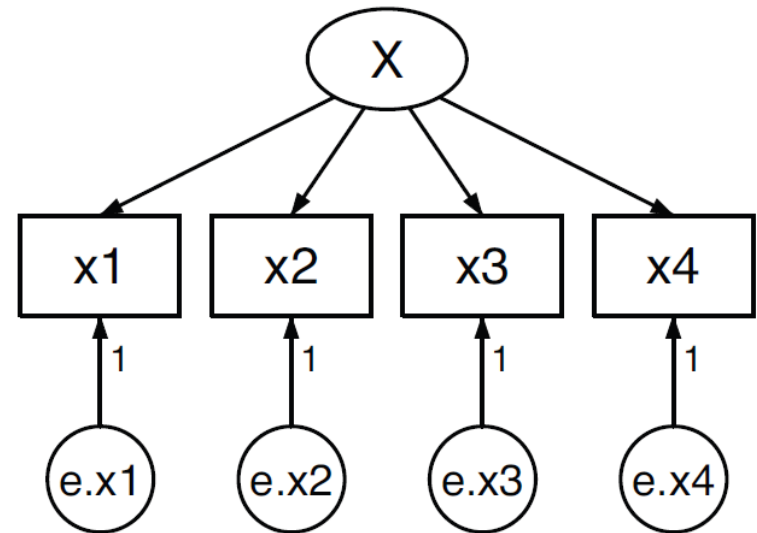
# Jargon

- Observed and Latent variables
- Paths and Covariance
- Endogenous and Exogenous variables
- Recursive and Nonrecursive models



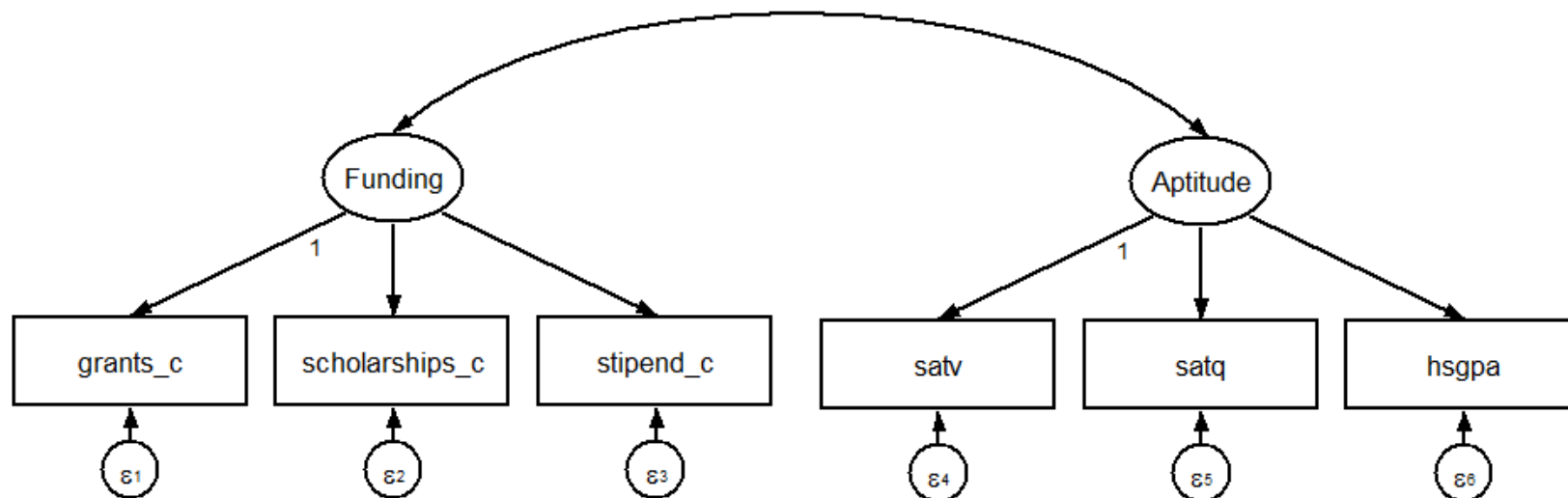
# Observed and Latent Variables

- **Observed variables** are variables that are included in our dataset. They are represented by rectangles. The variables  $x_1$ ,  $x_2$ ,  $x_3$  and  $x_4$  are observed variables in this path diagram.
- **Latent variables** are unobserved variables that we wish we had observed. They can be thought of as a composite score of other variables. They are represented by ovals. The variable  $X$  is a latent variable in this path diagram.



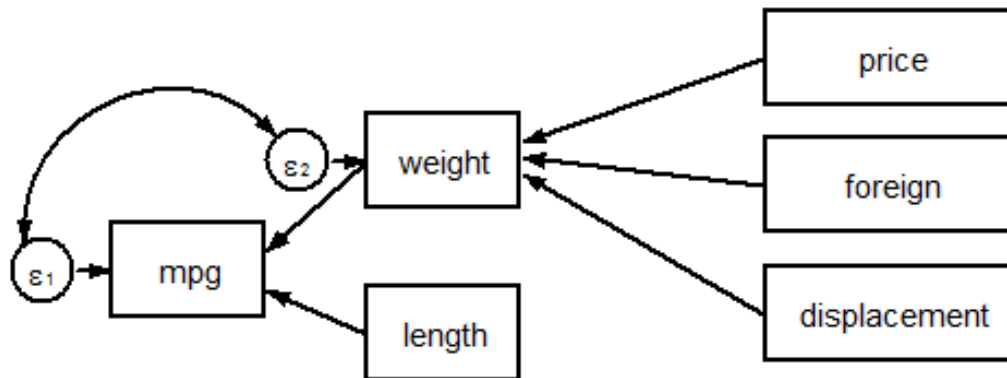
# Paths and Covariance

- **Paths** are direct relationships between variables. Estimated path coefficients are analogous to regression coefficients. They are represented by straight arrows.
- **Covariance** specify that two latent variables or error terms covary. They are represented by curved arrows.



# Exogenous and Endogenous Variables

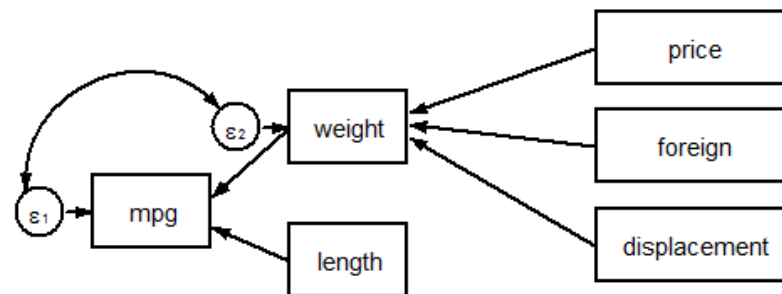
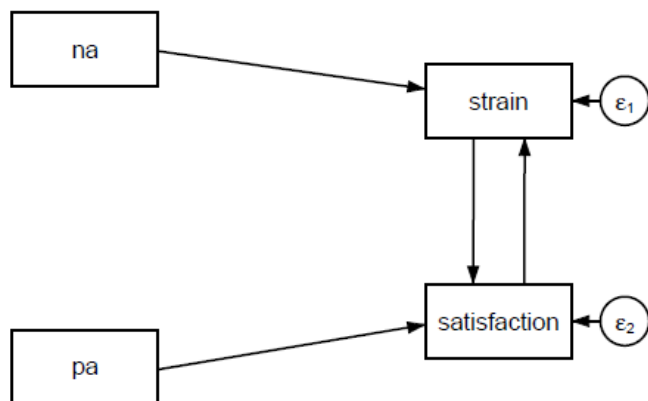
- **Exogenous** variables are determined outside the system of equations. There are no paths pointing to it. The variables `price`, `foreign`, `displacement` and `length` are exogenous.
- **Endogenous** variables are determined by the system of equations. At least one path points to it. The variables `weight` and `mpg` are endogenous.



- **Observed Exogenous**: a variable in a dataset that is treated as exogenous in the model
- **Latent Exogenous**: an unobserved variable that is treated as exogenous in the model.
- **Observed Endogenous**: a variable in a dataset that is treated as endogenous in the model
- **Latent Endogenous**: an unobserved variable that is treated as endogenous in the model.

# Recursive and Nonrecursive Systems

- **Recursive** models do not have any feedback loops or correlated errors.
- **Nonrecursive** models have feedback loops or correlated errors. These models have paths in both directions between one or more pairs of endogenous variables



# Notation

- Observed endogenous:  $\mathbf{y}$
- Observed exogenous:  $\mathbf{x}$
- Latent endogenous:  $\boldsymbol{\eta}$
- Latent exogenous:  $\boldsymbol{\xi}$
- Error of observed endogenous:  $\mathbf{e.y}$
- Error of latent endogenous:  $\mathbf{e.\eta}$
- All endogenous:  $\mathbf{Y} = \mathbf{y} \boldsymbol{\eta}$
- All exogenous:  $\mathbf{X} = \mathbf{x} \boldsymbol{\xi}$
- All error:  $\mathbf{e} = \mathbf{e.y} \mathbf{e.\eta}$

$$Y = BY + \Gamma X + \alpha + \zeta$$

We estimate:

- The coefficients  $\mathbf{B}$  and  $\mathbf{\Gamma}$
- The intercepts,  $\alpha$
- The means of the exogenous variables  $\boldsymbol{\kappa} = E(\mathbf{X})$
- The variances and covariances of the exogenous variables,  $\boldsymbol{\Phi} = \text{Var}(\mathbf{X})$
- The variances and covariances of the errors  $\boldsymbol{\Psi} = \text{Var}(\zeta)$

# Assumptions

- Large Sample Size
- Multivariate Normality
- Correct Model Specification



# Assumptions

- Large Sample Size
  - ML estimation relies on asymptotics, and large sample sizes are needed to obtain reliable parameter estimates.
  - Different suggestions regarding appropriate sample size have been given by different authors.
  - A common rule of thumb is to have a sample size of more than 200, although sometimes 100 is seen as adequate.
  - Other authors propose sample sizes relative to the number of parameters being estimated. Ratios of observations to free parameters from 5:1 up to 20:1 have been proposed.

# Assumptions

- Multivariate Normality
  - The likelihood that is maximized when fitting structural equation models using ML is derived under the assumption that the observed variables follow a multivariate normal distribution.
  - The assumption of multivariate normality can often be relaxed, particularly for exogenous variables.

# Assumptions

- Correct Model Specification
  - SEM assumes that no relevant variables are omitted from any equation in the model.
  - Omitted variable bias can arise in linear regression if an independent variable is omitted from the model and the omitted variable is correlated with other independent variables.
  - When fitting structural equation models with ML and all equations are fit jointly, errors can occur in equations other than the one with the omitted variable.

# What is Structural Equation Modeling?

- Brief history
- Path diagrams
- Key concepts, jargon and assumptions
- **Assessing model fit**
- The process of SEM

# Assessing Model Goodness of Fit

- Model Definitions
  - The **Saturated Model** assumes that all variables are correlated.
  - The **Baseline Model** assumes that no variables are correlated (except for exogenous variables when endogenous variables are present).
  - The **Specified Model** is the model that we fit

Likelihood Ratio  $\chi^2$  (baseline vs saturated models)

$$\chi_{bs}^2 = 2\{\log L_s - \log L_b\}$$

Likelihood Ratio  $\chi^2$  (specified vs saturated models)

$$\chi_{ms}^2 = 2\{\log L_s - \log L_m\}$$

where:

$L_b$  is the loglikelihood for the baseline model

$L_s$  is the loglikelihood for the saturated model

$L_m$  is the loglikelihood for the specified model

$$df_{bs} = df_s - df_b$$

$$df_{ms} = df_s - df_m$$

# Assessing Model Goodness of Fit

- Likelihood Ratio Chi-squared Test ( $\chi^2_{ms}$ )
- Akaike's Information Criterion (AIC)
- Swartz's Bayesian Information Criterion (BIC)
- Coefficient of Determination ( $R^2$ )
- Root Mean Square Error of Approximation (RMSEA)
- Comparative Fit Index (CFI)
- Tucker-Lewis Index (TLI)
- Standardized Root Mean Square Residual (SRMR)

# Assessing Model Goodness of Fit

Likelihood Ratio  $\chi^2$  (baseline vs saturated models)

$$\chi_{bs}^2 = 2\{\log L_s - \log L_b\}$$

where:

$L_s$  is the loglikelihood for the saturated model

$L_m$  is the loglikelihood for the specified model

$$df_{ms} = df_s - df_m$$

## Good fit indicated by:

- p-value > 0.05



# Assessing Model Goodness of Fit

Akaike's Information Criterion (AIC)

$$AIC = -2 \log L_m + 2df_m$$

Swartz's Bayesian Information Criterion (BIC)

$$BIC = -2 \log L_m + Ndf_m$$

**Good fit indicated by:**

- Used for comparing two models
- Smaller (in absolute value) is better

# Assessing Model Goodness of Fit

Coefficient of Determination ( $R^2$ )

$$R^2 = 1 - \frac{\det(\hat{\Psi})}{\det(\hat{\Sigma})}$$

**Good fit indicated by:**

- Values closer to 1 indicate good fit

# Assessing Model Goodness of Fit

- Root Mean Square Error of Approximation
  - Compares the current model with the saturated model
  - The null hypothesis is that the model fits

$$RMSEA = \sqrt{\frac{(\chi_{ms}^2 - df_{ms})}{(N - 1)df_{ms}}}$$

## Good fit indicated by:

- Hu and Bentler (1999):  $RMSEA < 0.06$
- Browne and Cudeck (1993)
  - Good Fit ( $RMSEA < 0.05$ )
  - Adequate Fit ( $RMSEA$  between 0.05 and 0.08)
  - Poor Fit ( $RMSEA > 0.1$ )
- P-value  $> 0.05$

# Assessing Model Goodness of Fit

- Comparative Fit Index (CFI)
  - Compares the current model with the baseline model

$$CFI = 1 - \frac{\chi_{ms}^2 - df_{ms}}{\chi_{bs}^2 - df_{bs}}$$

## Good fit indicated by:

- $CFI > 0.95$  (sometimes 0.90)

# Assessing Model Goodness of Fit

## Tucker-Lewis Index (TLI)

- Compares the current model with the baseline model

$$TLI = 1 - \frac{(\chi_{bs}^2 / df_{bs}) - (\chi_{bs}^2 / df_{bs})}{(\chi_{bs}^2 / df_{bs}) - 1}$$

### Good fit indicated by:

- $TLI > 0.95$

# Assessing Model Goodness of Fit

## Standardized Root Mean Square Residual (SRMR)

- SRMR is a measure of the average difference between the observed and model implied correlations. This will be close to 0 when the model fits well. Hu and Bentler (1999) suggest values close to .08 or below.

### Good fit indicated by:

- $SRMR < 0.08$

# The Process of SEM

- Specify the model
- Fit the model
- Evaluate the model
- Modify the model
- Interpret and report the results

# Outline

- Introduction to Stata
- What is structural equation modeling?
- **Structural equation modeling in Stata**
- Continuous outcome models using **sem**
- Multilevel generalized models using **gsem**
- Demonstrations and Questions



# Structural Equation Modeling in Stata

- Getting your data into Stata
- The SEM Builder
- The **sem** syntax
- The **gsem** syntax
- Differences between **sem** and **gsem**

# Getting Data Into Stata

- Can import data using
  - `insheet`
  - `infile`
  - `import excel`
- Can open observation level data with `use`
- Can open summary data with `ssd`

# Getting Data Into Stata

```
clear
```

```
ssd init fygpa grants scholarships stipend
```

```
ssd set obs 100
```

```
ssd set means 2.40 6.43 5.34 0.85
```

```
ssd set cov 0.53 \ ///  
            -0.21 90.99 \ ///  
            0.72 -8.98 93.29 \ ///  
            0.06 4.01 0.25 1.54
```

Note that we will not be able to use **gsem** with summary data

# Getting Data Into Stata

```
. ssd list
```

```
Observations = 100
```

```
Means:
```

fygpa	grants	scholarships	stipend
2.4	6.43	5.34	.85

```
Variances implicitly defined; they are the diagonal of  
the covariance matrix.
```

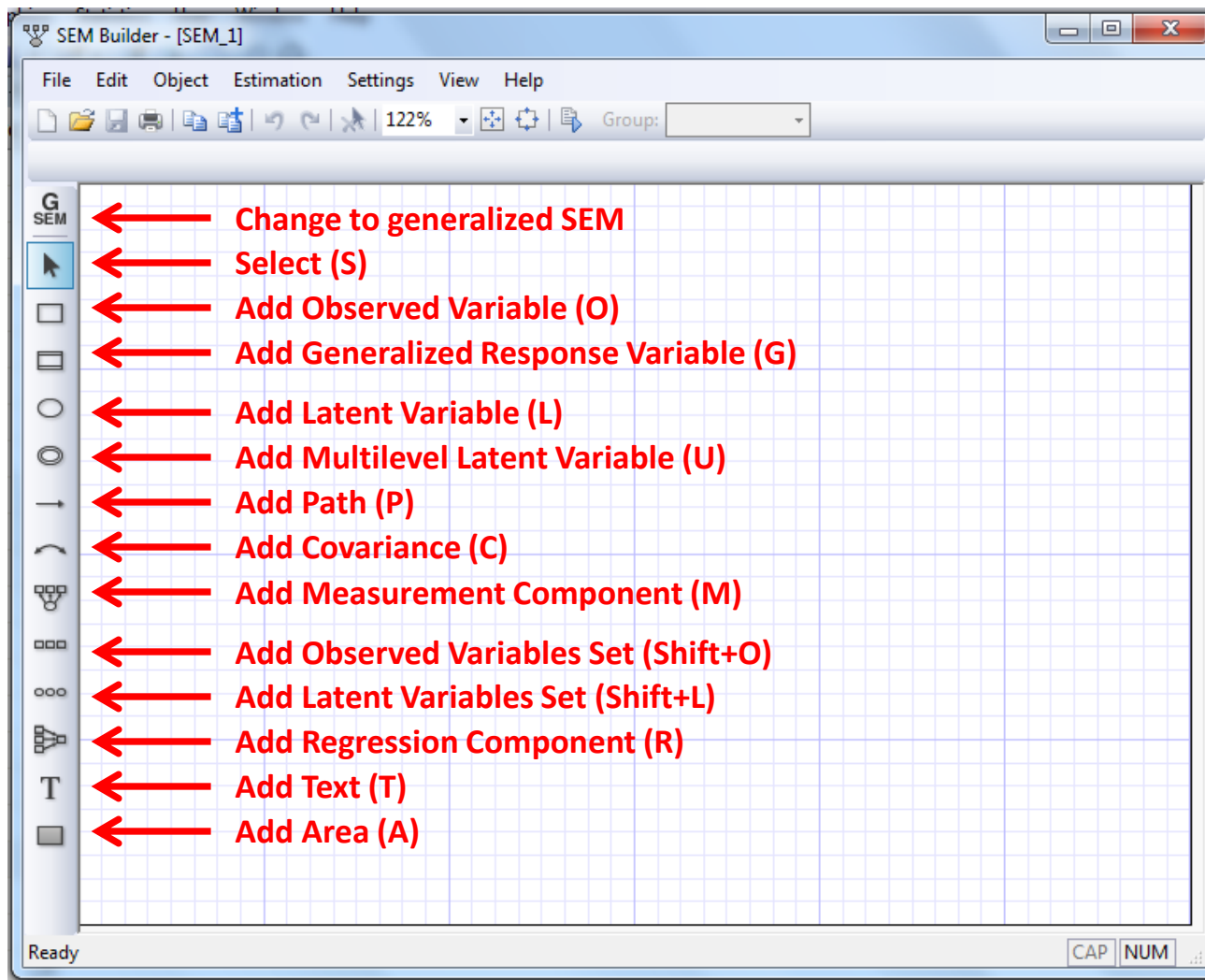
```
Covariances:
```

fygpa	grants	scholarships	stipend
.53			
-.21	90.99		
.72	-8.98	93.29	
.06	4.01	.25	1.54

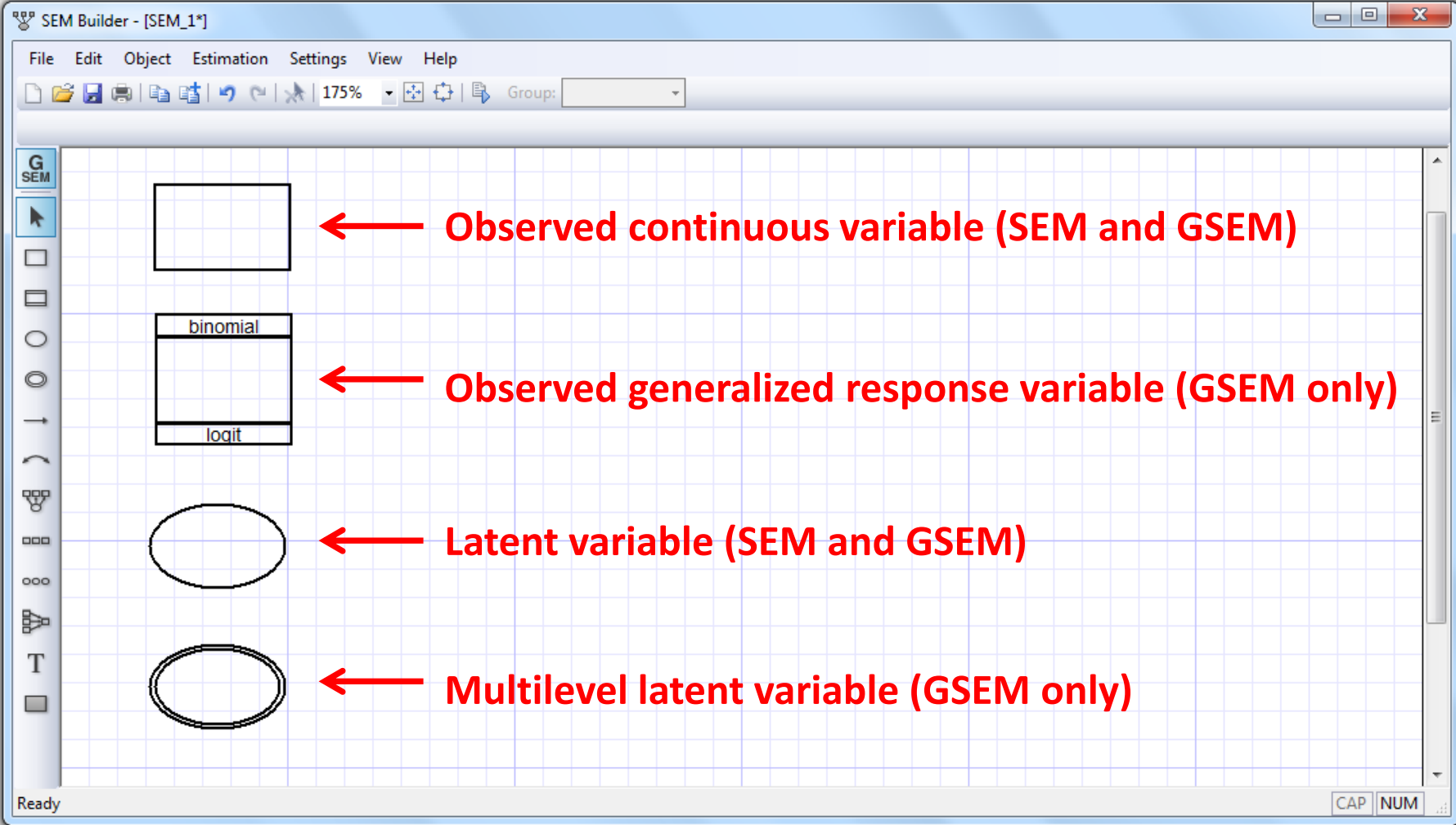
# Structural Equation Modeling in Stata

- Getting your data into Stata
- **The SEM Builder**
- The **sem** syntax
- The **gsem** syntax
- Differences between **sem** and **gsem**

# We can draw path diagrams using Stata's SEM Builder



# Drawing variables in Stata's SEM Builder



The screenshot shows the Stata SEM Builder window with a grid background. On the left is a toolbar with icons for drawing variables. Four red arrows point from text labels to specific icons in the toolbar:

- ← **Observed continuous variable (SEM and GSEM)**: Points to a simple rectangle icon.
- ← **Observed generalized response variable (GSEM only)**: Points to a rectangle icon with 'binomial' in the top section and 'logit' in the bottom section.
- ← **Latent variable (SEM and GSEM)**: Points to an oval icon.
- ← **Multilevel latent variable (GSEM only)**: Points to a double-lined oval icon.

The window title is 'SEM Builder - [SEM\_1\*]' and the status bar at the bottom shows 'Ready' and 'CAP NUM'.

# We can draw path diagrams using Stata's SEM Builder

The screenshot shows the Stata SEM Builder interface with a path diagram. The diagram illustrates a structural equation model with the following components:

- Latent Variables:** 'Funding' and 'Aptitude' (represented by ovals).
- Observed Variables:** 'grants2', 'scholarships2', 'stipend2', 'satv', 'satq', 'hsgpa', 'fygpa', 'instate', 'male', 'credithrs', and 'ptindex' (represented by rectangles).
- Error Terms:**  $\epsilon_1$ ,  $\epsilon_2$ ,  $\epsilon_3$ , and  $\epsilon_4$  (represented by circles).
- Path Diagram:**
  - 'Funding' points to 'grants2', 'scholarships2', and 'stipend2'.
  - 'Aptitude' points to 'satv', 'satq', and 'hsgpa'.
  - 'Funding' and 'Aptitude' both point to 'fygpa'.
  - 'university<sub>1</sub>' (a double-lined oval) points to 'fygpa'.
  - 'fygpa' points to 'instate', 'male', 'credithrs', and 'ptindex'.
  - 'satv' points to 'grants2'.
  - 'satq' points to 'scholarships2'.
  - 'hsgpa' points to 'stipend2'.
  - 'fygpa' has an associated error term  $\epsilon_4$ .

The right-hand side of the window displays the 'Details' panel for the selected variable 'fygpa', showing the following statistics:

Intercept	
_cons	.8974062
SE	.0967615
z	9.27441
P	1.79e-20
CI-LB	.707757
CI-UB	1.087055
Std. intercept	
_cons	1.2396
SE	.1350467
z	9.179051
P	4.35e-20
CI-LB	.9749137
CI-UB	1.504287



# Structural Equation Modeling in Stata

- Getting your data into Stata
- The SEM Builder
- **The `sem` syntax**
- The `gsem` syntax
- Differences between `sem` and `gsem`

# sem syntax

```
sem paths [if] [in] [weight] [, options]
```

- Paths are specified in parentheses and correspond to the arrows in the path diagrams we saw previously.
- Arrows can point in either direction.
- Paths can be specified individually, or multiple paths can be specified within a single set of parentheses.

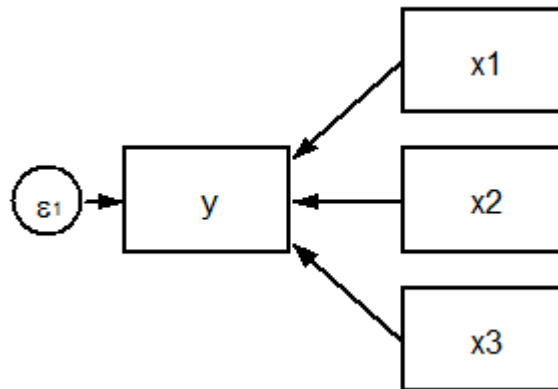
# sem syntax examples

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

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

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

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

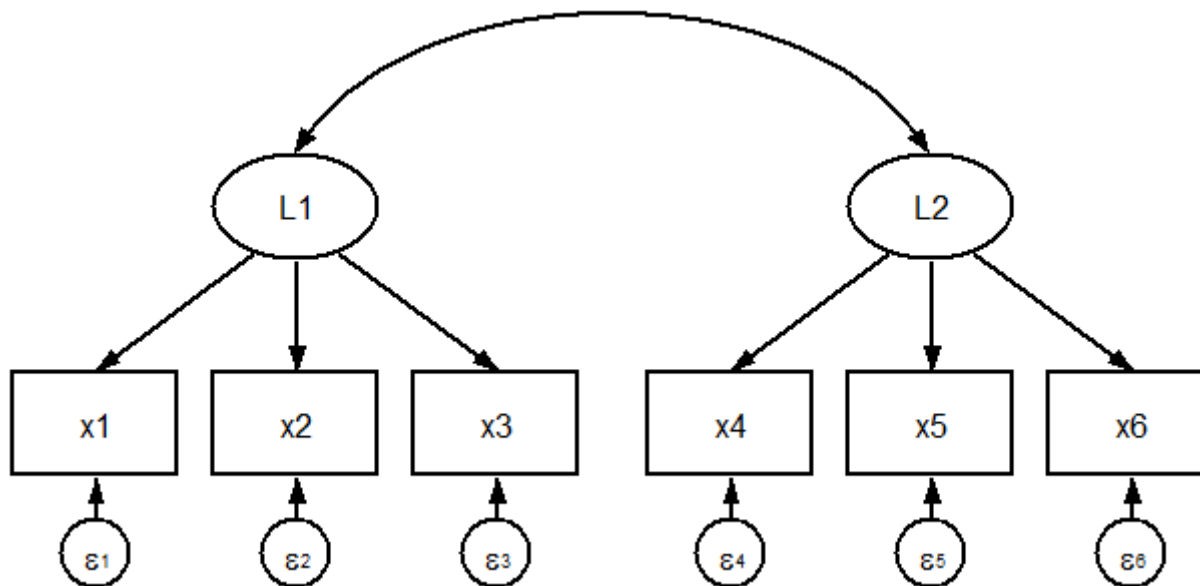


# sem syntax examples

```
sem (L1 <- x1 x2 x3) (L2 <- x4 x5 x6)
```

```
sem (x1 x2 x3 -> L1) (x1 x2 x3 -> L1)
```

```
sem (L1 <- x1) (L1 <- x2) (L1 <- x3) ///  
    (L2 <- x4) (L2 <- x5) (L2 <- x6)
```



# sem syntax examples

```
sem (L1 <- x1 x2 x3) (L2 <- x4 x5 x6), standardized
```

```
sem (L1 <- x1@1 x2 x3) (L2 <- x4@1 x5 x6)
```

```
sem (L1 <- x1@a x2 x3) (L2 <- x4@a x5 x6)
```

```
sem (latent1 <- x1 x2 x3) (latent2 <- x4 x5 x6), ///  
latent(latent1 latent2) nocapslatent
```

```
sem (L1 <- x1 x2 x3) (L2 <- x4 x5 x6), group(female)
```

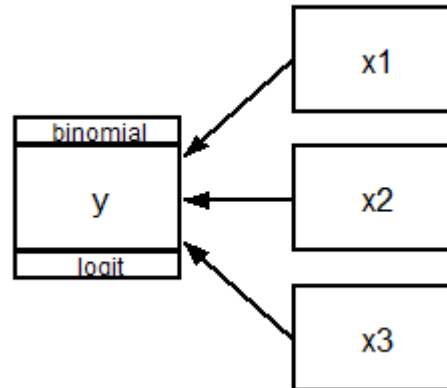
# Structural Equation Modeling in Stata

- Getting your data into Stata
- The SEM Builder
- The `sem` syntax
- **The `gsem` syntax**
- Differences between `sem` and `gsem`

# gsem syntax examples

```
gsem (y <- x1 x2 x3, family(bernoulli) link(logit))
```

```
gsem (y <- x1 x2 x3), logit
```



# Families and Link Functions

	identity	log	logit	probit	cloglog
gaussian	X	X			
gamma		X			
bernoulli			X	X	X
binomial			X	X	X
ordinal			X	X	X
multinomial			X		
Poisson		X			
nbinomial		X			

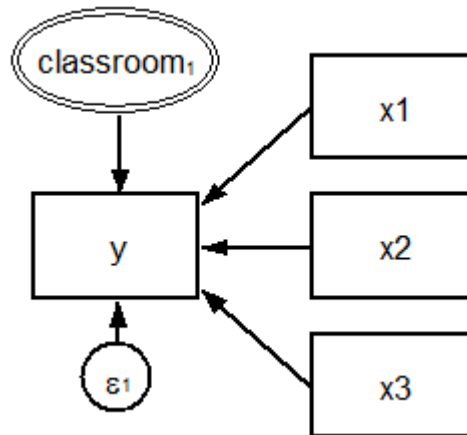


# Families and Link Functions

Option	Synonym for:
<code>cloglog</code>	<code>family(bernoulli) link(cloglog)</code>
<code>gamma</code>	<code>family(gamma) link(log)</code>
<code>logit</code>	<code>family(bernoulli) link(logit)</code>
<code>nbreg</code>	<code>family(nbinomial mean) link(log)</code>
<code>mlogit</code>	<code>family(multinomial) link(logit)</code>
<code>ocloglog</code>	<code>family(ordinal) link(cloglog)</code>
<code>ologit</code>	<code>family(ordinal) link(logit)</code>
<code>oprobit</code>	<code>family(ordinal) link(probit)</code>
<code>poisson</code>	<code>family(poisson) link(log)</code>
<code>probit</code>	<code>family(bernoulli) link(probit)</code>
<code>regress</code>	<code>family(gaussian) link(identity)</code>

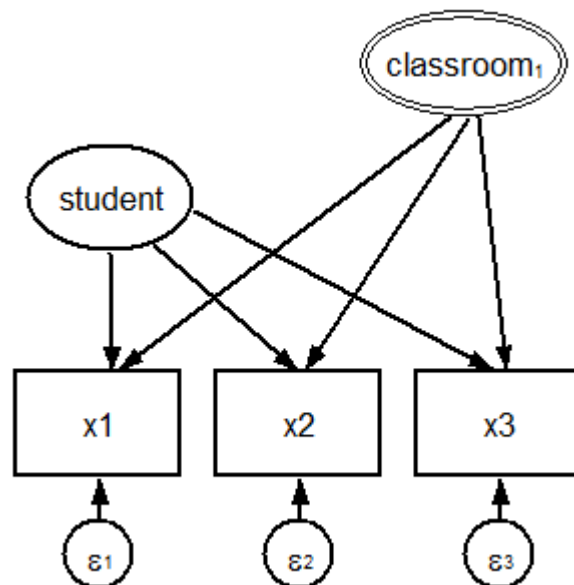
# gsem syntax examples

```
gsem (y <- x1 x2 x3)          ///  
     (y <- M1[classroom]),    ///  
latent(M1) nocapslatent
```



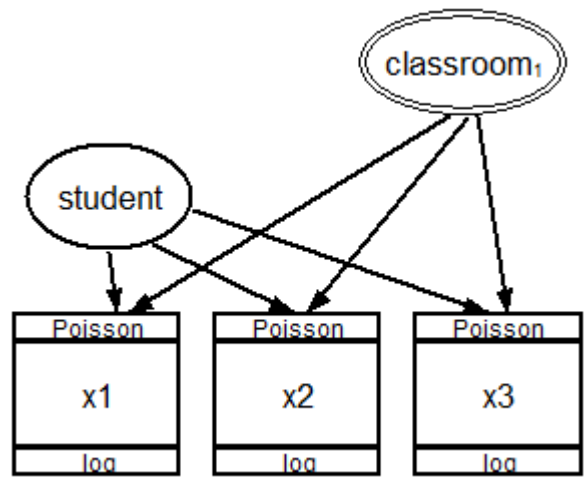
# gsem syntax examples

```
gsem (M1[classroom] -> x1 x2 x3)          ///  
     (student -> x1 x2 x3) ,              ///  
     latent(student M1 ) nocapslatent
```



# gsem syntax examples

```
gsem (M1[classroom] -> x1 x2 x3, family(poisson) link(log)) ///
(student -> x1 x2 x3, family(poisson) link(log)),          ///
latent(student M1 ) nocapslatent
```



# Structural Equation Modeling in Stata

- Getting your data into Stata
- The SEM Builder
- The **sem** syntax
- The **gsem** syntax
- **Differences between sem and gsem**

# Differences Between **sem** and **gsem**

- **sem** features not available with **gsem**:
  - Estimation methods MLMV and ADF
  - Fitting models with summary statistics data (SSD)
  - Specialized syntax for multiple-group models
  - Estimates adjusted for complex survey design
  - estat commands for goodness of fit, indirect effects, modification indices, and covariance residuals

# Differences Between **sem** and **gsem**

- **gsem** features not available with **sem**:
  - Generalized-linear response variables
  - Multilevel models
  - Factor-variable notation may be used
  - Equation-wise deletion of observations with missing values
  - margins, contrast, and pwcompare command may be used after gsem

# Differences Between **sem** and **gsem**

- You may obtain different likelihood values when fitting the same model with **sem** and **gsem**.
  - The likelihood for **sem** is derived including estimation of the means, variances, and covariances of the observed exogenous variables.
  - The likelihood for the model fit by **gsem** is derived as conditional on the values of the observed exogenous variables.
  - Normality of observed exogenous variables is never assumed with **gsem**.



# Outline

- Introduction to Stata
- What is structural equation modeling?
- Structural equation modeling in Stata
- **Continuous outcome models using `sem`**
- Multilevel generalized models using **`gsem`**
- Demonstrations and Questions

# Continuous Outcome Models Using **sem**

- Example Data
- Means
- Correlation
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Example Data

```
. use cair.dta, clear
(Example data for the California Association for Institutional Research Workshop)
```

```
. describe
```

storage	display	value		
variable name	type	format	label	variable label
id	int	%9.0g		Identification Number
university	byte	%9.0g		University ID
college	byte	%11.0g	college	Primary college of major
private	byte	%9.0g	private	Private or public university?
fygpa	double	%4.2f		First-year GPA
ret_yr1	byte	%8.0g	YesNo	* First-year retention
instate	byte	%12.0g	instate	* In state residency
male	byte	%8.0g	male	Male
greek	byte	%8.0g	YesNo	* Member of a Greek society
withdrawn	double	%3.0f		* Credits withdrawn or incomplete
credithrs	double	%3.0f		* Average number of credits hours attempted per term
ptindex	double	%3.0f		* % courses taken in 1st year from part time faculty
grants	double	%5.1f		* Grant money (x1000 dollars)
scholarships	double	%5.1f		* Scholarship money (x1000 dollars)
stipend	double	%5.1f		* Student work income (x1000 dollars)
				* indicated variables have notes

Sorted by: id

# Example Data

**. summarize**

Variable	Obs	Mean	Std. Dev.	Min	Max
id	12958	6479.5	3740.797	1	12958
university	12958	10.45956	5.735442	1	20
college	12958	3.052091	1.495687	1	5
private	12958	.4972218	.5000116	0	1
fygpa	12875	2.398844	.7274577	0	4
ret_yr1	12958	.8924217	.3098591	0	1
instate	12958	.730977	.4434691	0	1
male	12958	.4069301	.4912806	0	1
greek	12958	.2218707	.4155206	0	1
withdrawn	12947	3.864951	10.26619	0	100
credithrs	12947	15.62393	1.025208	9	24
ptindex	12947	44.0851	18.11552	0	100
grants	12958	6.399958	9.520231	0	49.558
scholarships	12958	5.319597	9.637058	0	69.288
stipend	12958	.8426065	1.237821	0	10.79976

# Example Data

```
. notes _dta
```

```
_dta:
```

1. Data from Bryce Mason at UC Riverside
2. Data set of new freshmen (starting college) across a number of years at a mid-sized, private, moderately selective university
3. It focuses only on the first year of enrollment and first-year retention (or GPA) as the outcome of interest.

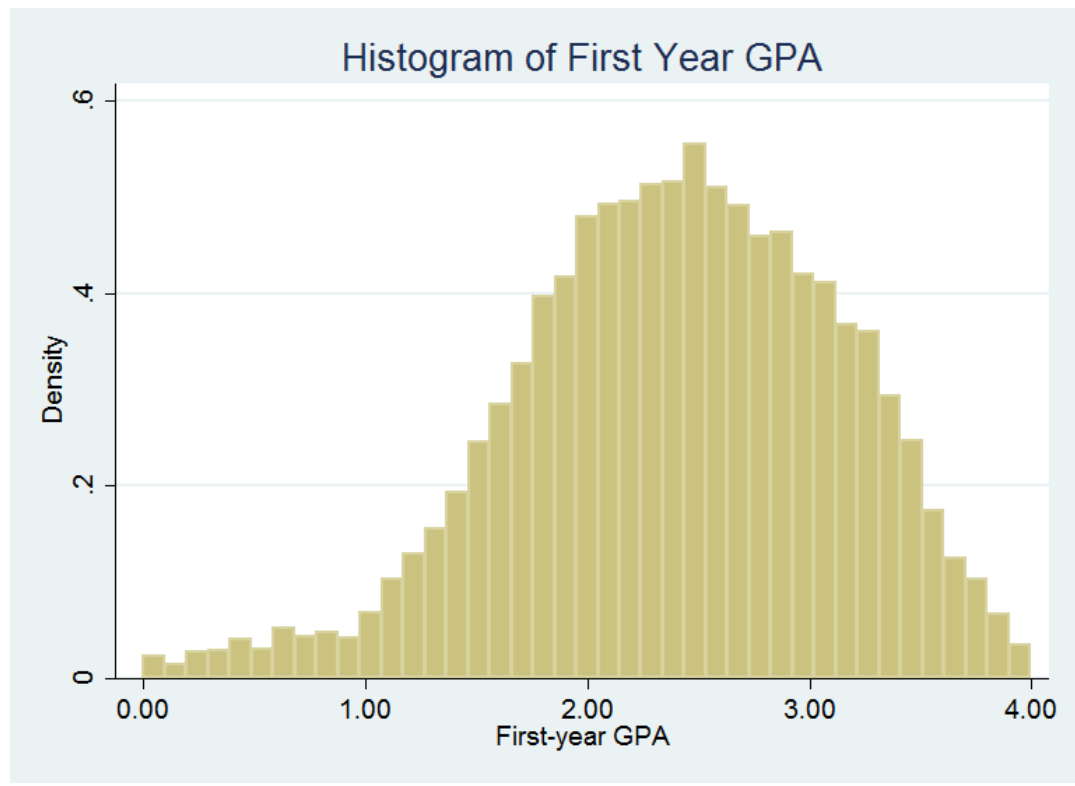
```
. notes ret_yr1
```

```
ret_yr1:
```

1. So-called first-year retention. Measures whether the student was enrolled in the fall term of what would have been their second year of studies

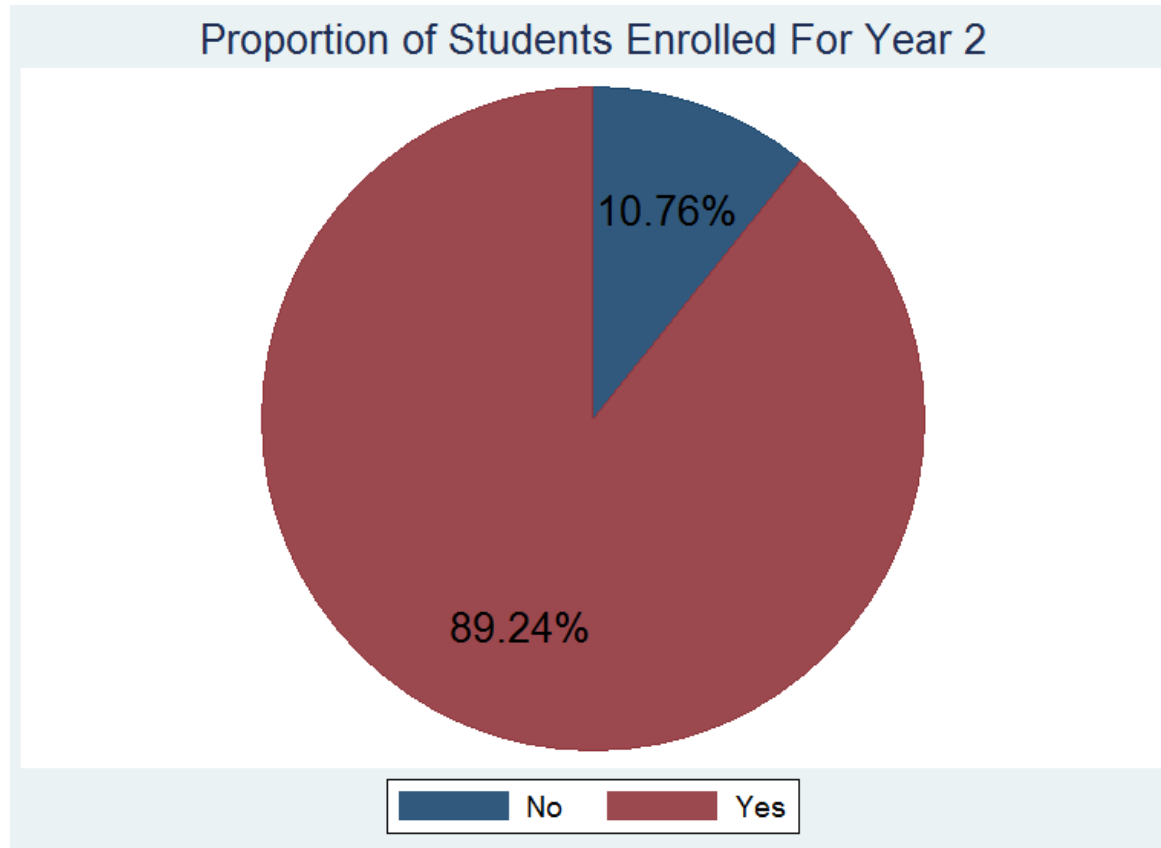
# Example Data

```
histogram fygpa, title(Histogram of First Year GPA)
```



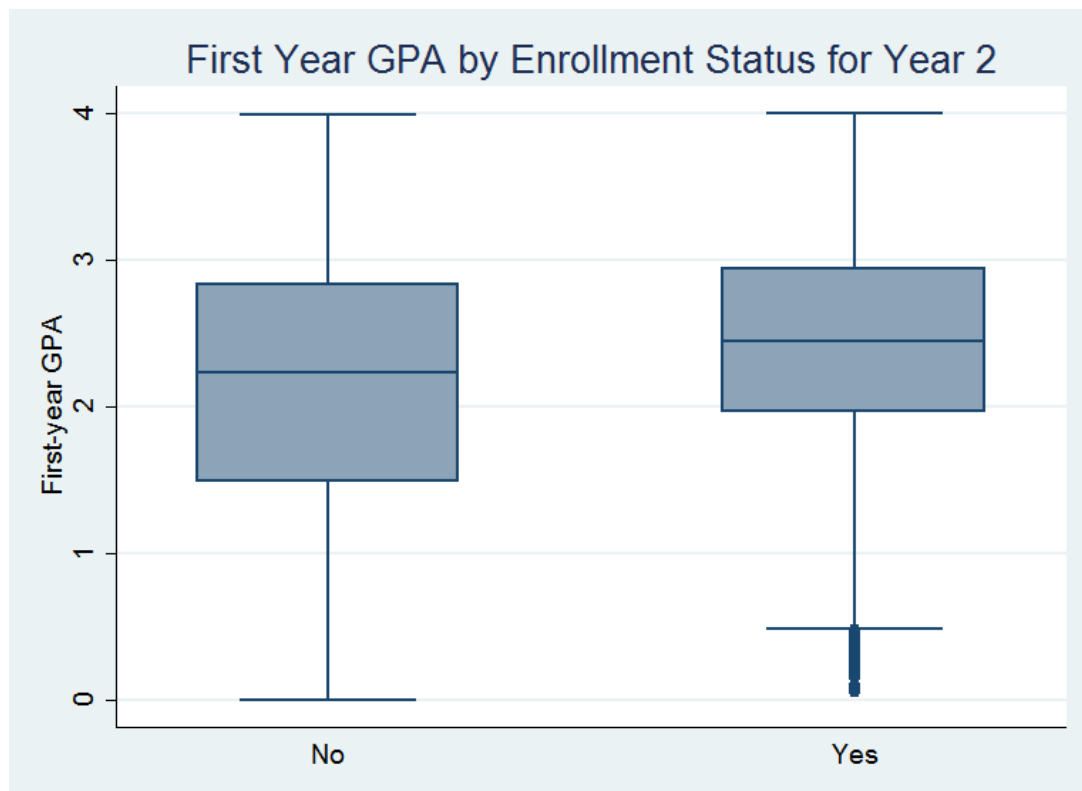
# Example Data

```
graph pie, over(ret_yr1)           ///  
    plabel(_all percent, size(large))  ///  
    title(Proportion of Students Enrolled For Year 2)
```



# Example Data

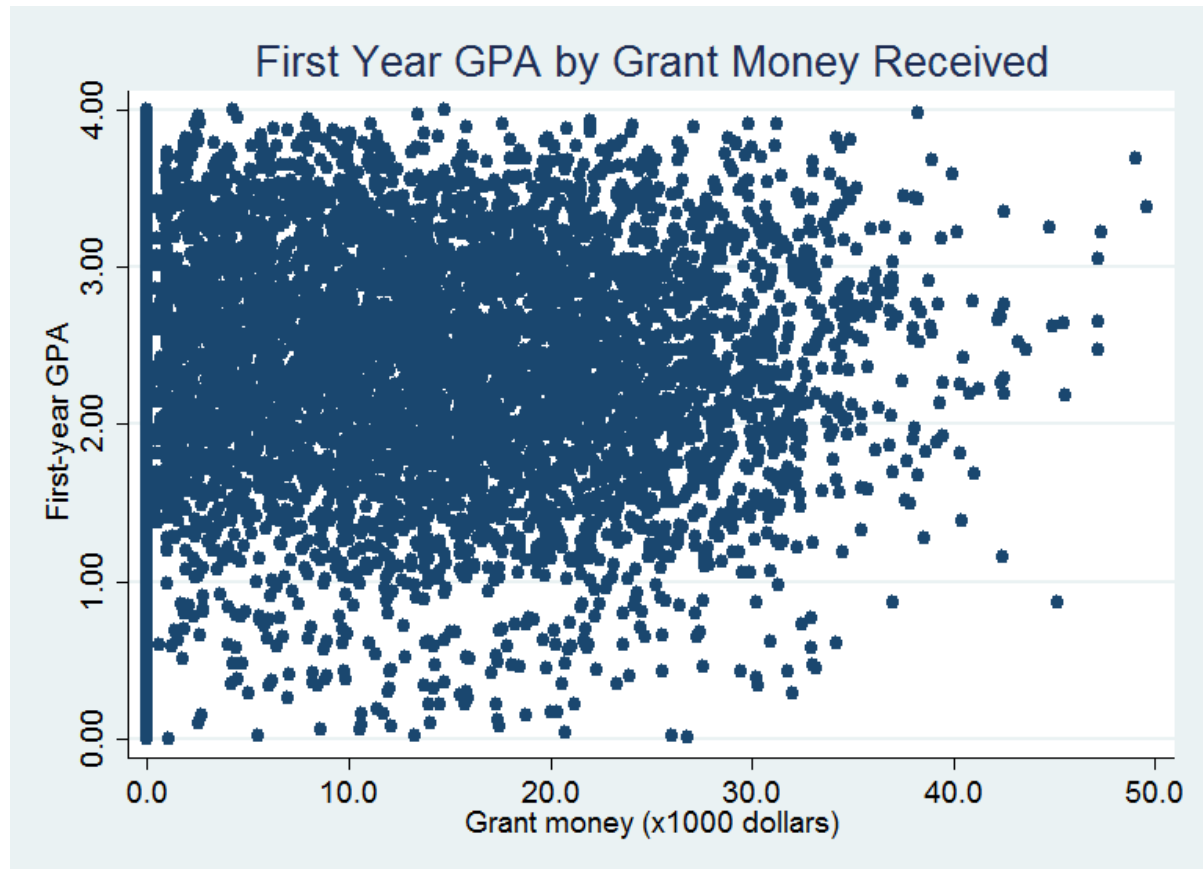
```
graph box fygpa, over(ret_yr1)      ///  
    title(First Year GPA by Enrollment Status for Year 2)
```





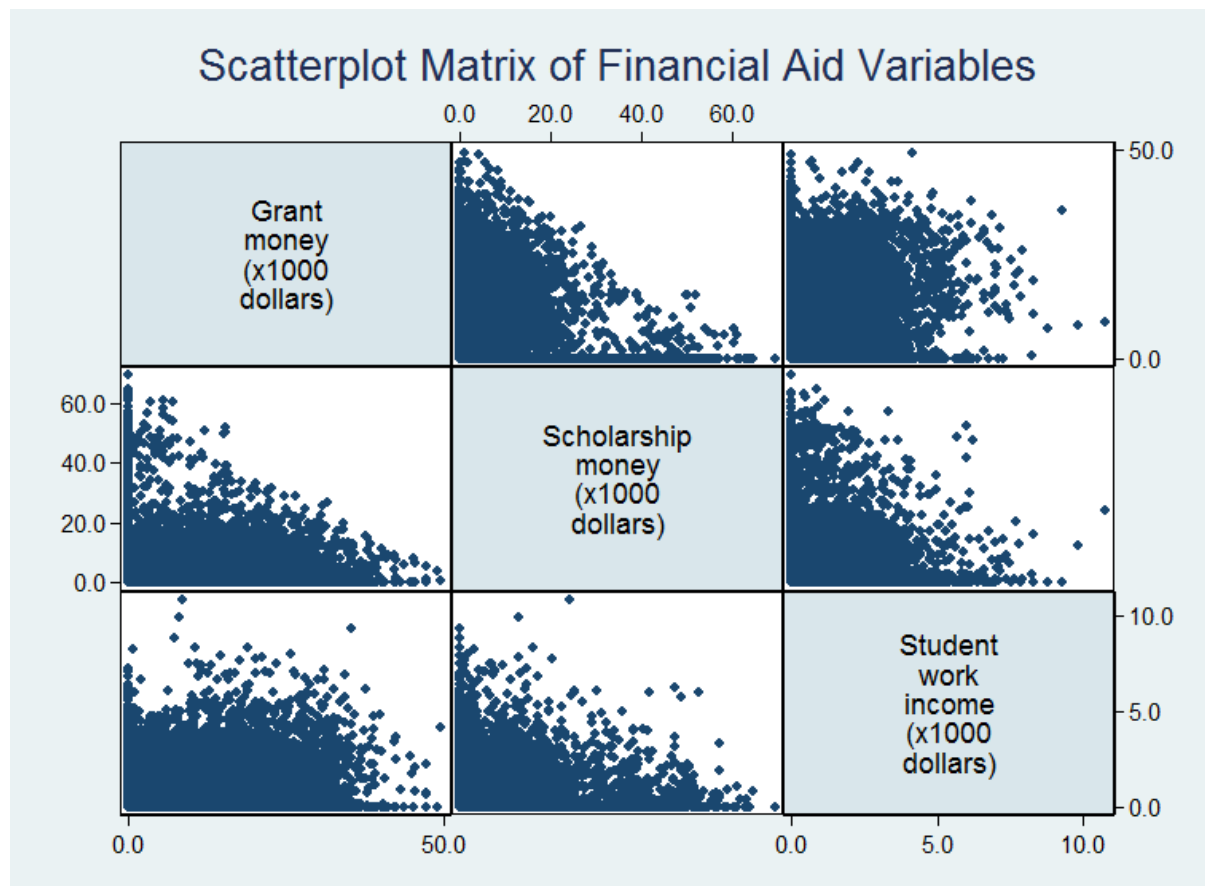
# Example Data

```
twoway (scatter fygpa grants),      ///  
       title(First Year GPA by Grant Money Received)
```



# Example Data

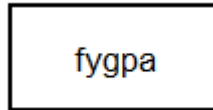
```
graph matrix grants scholarships stipend, ///  
title(Scatterplot Matrix of Financial Aid Variables)
```



# Continuous Outcome Models Using **sem**

- Example Data
- **Means**
- Correlation
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Sample Mean Path Diagram



# Sample Mean Syntax

Syntax using **mean** :

```
mean fygpa
```

Syntax using **sem** :

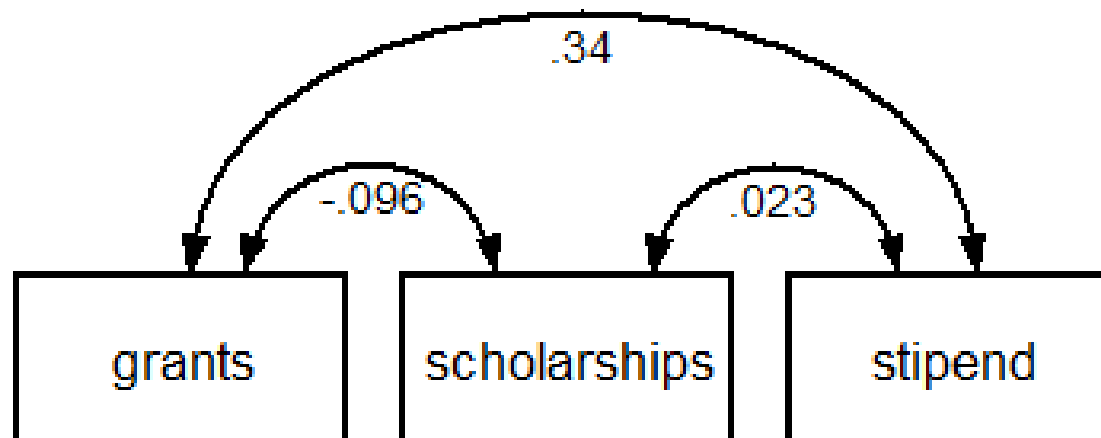
```
sem fygpa
```



# Continuous Outcome Models Using `sem`

- Example Data
- Means
- **Correlation**
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Correlation Path Diagram





# Correlation Syntax

Syntax using **correlate**:

```
correlate grants scholarships stipend
```

Syntax using **sem**:

```
sem grants scholarships stipend, standardized
```

# Correlation Results

Results using `correlate` :

	grants	scholarships	stipend
grants	1.0000		
scholarships	-0.0958	1.0000	
stipend	0.3402	0.0225	1.0000

Results using `sem` :

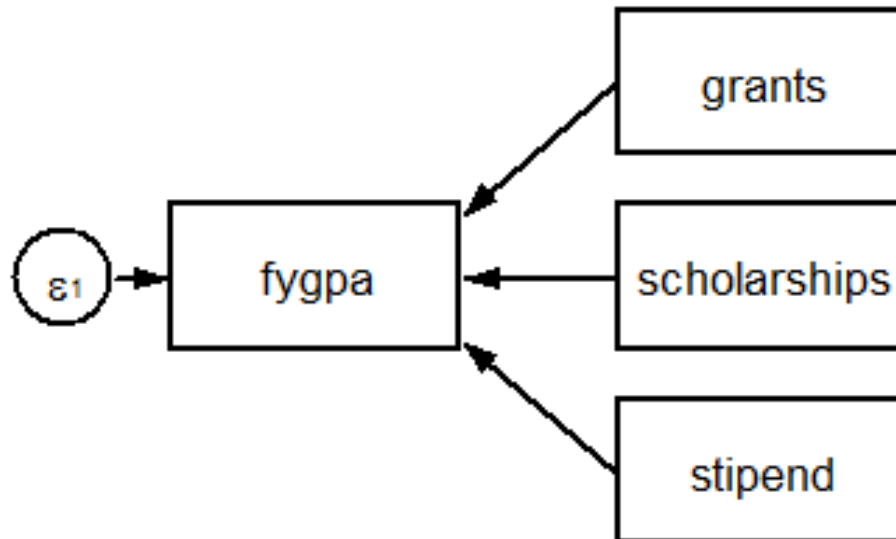
Standardized	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
mean(grants)	.6722741	.0097268	69.12	0.000	.6532098	.6913384
mean(scholarships)	.5520152	.0094303	58.54	0.000	.5335321	.5704982
mean(stipend)	.680744	.0097496	69.82	0.000	.6616353	.6998528
var(grants)	1	.			.	.
var(scholarships)	1	.			.	.
var(stipend)	1	.			.	.
cov(grants, scholarships)	<b>-.095848</b>	.0087041	-11.01	0.000	-.1129077	-.0787884
cov(grants, stipend)	.3402038	.007768	43.80	0.000	.3249787	.3554289
cov(scholarships, stipend)	.0225183	.0087803	2.56	0.010	.0053091	.0397274

LR test of model vs. saturated:  $\chi^2(0) = 0.00$ , Prob >  $\chi^2 = .$

# Continuous Outcome Models Using `sem`

- Example Data
- Means
- Correlation
- **Linear Regression**
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Linear Regression Path Diagram



# Linear Regression Syntax

Syntax using **regress** :

```
regress fygpa grants scholarships stipend
```

Syntax using **sem** :

```
sem fygpa <- grants scholarships stipend
```

# Linear Regression Results

Results using **regress** :

fygpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
grants	-.003563	.0007131	-5.00	0.000	-.0049608	-.0021651
scholarships	.0072665	.0006628	10.96	0.000	.0059673	.0085657
stipend	.04439	.0054608	8.13	0.000	.0336861	.055094
_cons	2.345305	.0090911	257.98	0.000	2.327485	2.363125

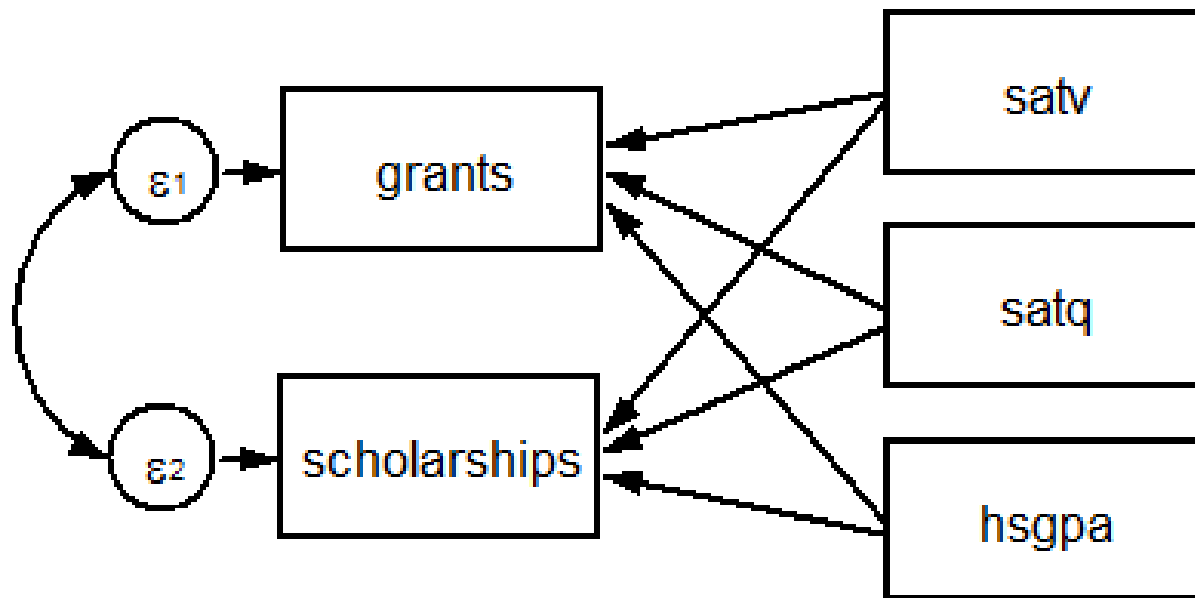
Results using **sem** :

	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural						
fygpa <-						
grants	-.003563	.000713	-5.00	0.000	-.0049605	-.0021654
scholarships	.0072665	.0006627	10.97	0.000	.0059676	.0085653
stipend	.04439	.0054599	8.13	0.000	.0336888	.0550913
_cons	2.345305	.0090897	258.02	0.000	2.327489	2.36312
var(e.fygpa)	.5206841	.0064896			.5081189	.5335601

# Continuous Outcome Models Using `sem`

- Example Data
- Means
- Correlation
- Linear Regression
- **Multivariate Regression**
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Multivariate Regression Path Diagram





# Multivariate Regression Syntax

## Syntax using `mvreg`:

```
mvreg grants scholarships = satv satq hsgpa
```

## Syntax using `sem`:

```
sem (grants scholarships <- satv satq hsgpa),    ///  
    cov(e.scholarships*e.grants)
```

```
sem (grants          <- satv satq hsgpa)        ///  
    (scholarships <- satv satq hsgpa),          ///  
    cov(e.scholarships*e.grants)
```

# Multivariate Regression Results

Results using `mvreg` :

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
<b>grants</b>						
satv	-.0556837	.0997707	-0.56	0.577	-.251249	.1398816
satq	-.0317529	.1027744	-0.31	0.757	-.2332059	.1697001
hsgpa	.0946835	.1025807	0.92	0.356	-.1063898	.2957568
_cons	6.467567	.1268666	50.98	0.000	6.218889	6.716244
<b>scholarships</b>						
satv	.2711446	.1009657	2.69	0.007	.0732369	.4690524
satq	.1581007	.1040054	1.52	0.129	-.0457652	.3619666
hsgpa	-.1293269	.1038093	-1.25	0.213	-.3328086	.0741548
_cons	4.975286	.1283862	38.75	0.000	4.72363	5.226942

# Multivariate Regression Results

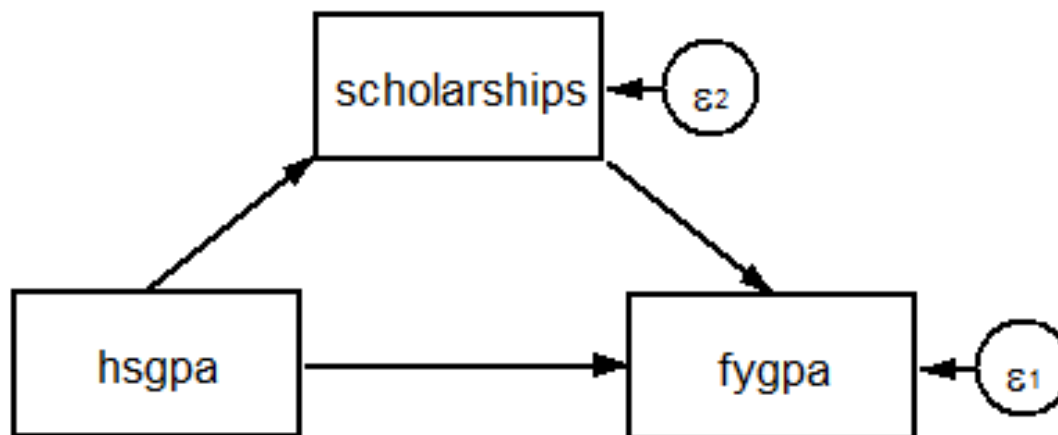
Results using `sem`:

	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
<b>Structural</b>						
grants <-						
satv	-.0556837	.0997552	-0.56	0.577	-.2512003	.1398329
satq	-.0317529	.1027584	-0.31	0.757	-.2331556	.1696499
hsgpa	.0946835	.1025647	0.92	0.356	-.1063397	.2957067
_cons	6.467567	.1268469	50.99	0.000	6.218951	6.716182
<b>scholar~s &lt;-</b>						
satv	.2711446	.10095	2.69	0.007	.0732862	.469003
satq	.1581007	.1039892	1.52	0.128	-.0457144	.3619158
hsgpa	-.1293269	.1037932	-1.25	0.213	-.3327578	.0741041
_cons	4.975286	.1283662	38.76	0.000	4.723693	5.226879
var(e.grants)	90.97287	1.133844			88.7775	93.22253
var(e.scholar~s)	93.1652	1.161168			90.91693	95.46908
cov(e.grants, e.scholarships)	-8.958597	.8151842	-10.99	0.000	-10.55633	-7.360866

# Continuous Outcome Models Using **sem**

- Example Data
- Means
- Correlation
- Linear Regression
- Multivariate Regression
- **Path Analysis and Mediation**
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Path Analysis Diagram



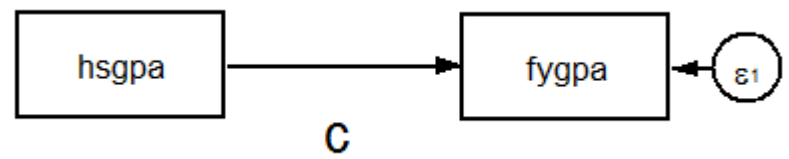
# Path Analysis Results

`sem (fygpa <- hsgpa scholarships) (scholarships <- hsgpa)`

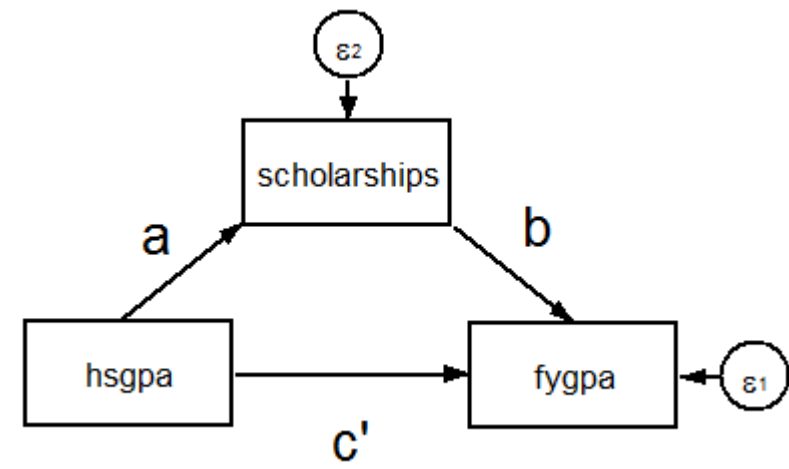
	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural						
fygpa <-						
scholarships	.0075858	.0006492	11.68	0.000	.0063134	.0088583
hsgpa	.1302288	.006181	21.07	0.000	.1181143	.1423433
_cons	2.280555	.0080434	283.53	0.000	2.26479	2.29632
scholarships <-						
hsgpa	.099537	.0839023	1.19	0.235	-.0649085	.2639826
_cons	5.283719	.0987621	53.50	0.000	5.090149	5.47729
var(e.fygpa)	.5061307	.0063082			.4939167	.5186467
var(e.scholarships)	93.27048	1.16248			91.01966	95.57695

# Mediation Analysis

**Total Effect** (c) of high school GPA on first year GPA



**Indirect Effect** (a & b) of high school GPA on first year GPA through the *mediator* scholarships



**Direct Effect** (c') of high school GPA on first year GPA

$$c = c' + ab$$

### estat teffects, compact

Direct effects

	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural						
fygpa <-						
scholarships	.0075858	.0006492	11.68	0.000	.0063134	.0088583
hsgpa	.1302288	.006181	21.07	0.000	.1181143	.1423433
scholars~s <-						
hsgpa	.099537	.0839023	1.19	0.235	-.0649085	.2639826

Indirect effects

	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural						
fygpa <-						
hsgpa	.0007551	.0006397	1.18	0.238	-.0004988	.0020089
scholars~s <-						

Total effects

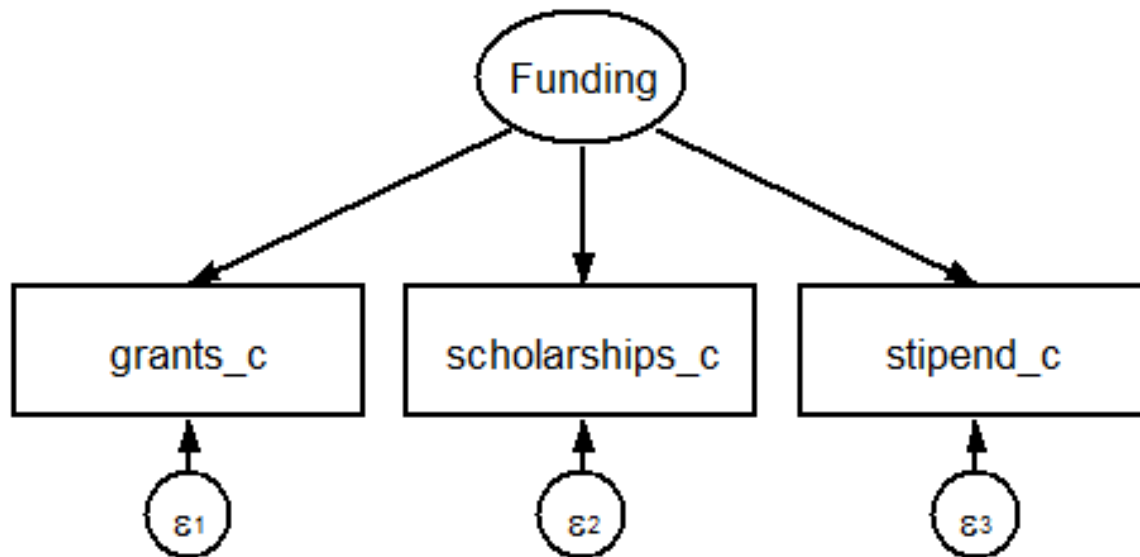
	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural						
fygpa <-						
scholarships	.0075858	.0006492	11.68	0.000	.0063134	.0088583
hsgpa	.1309839	.0062133	21.08	0.000	.118806	.1431618
scholars~s <-						
hsgpa	.099537	.0839023	1.19	0.235	-.0649085	.2639826



# Continuous Outcome Models Using `sem`

- Example Data
- Means
- Correlation
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- **Confirmatory Factor Analysis (CFA)**
- Structural Equation Models (SEM)
- Multi-group SEM
- SEM For Complex Survey Data

# Confirmatory Factor Analysis Path Diagram



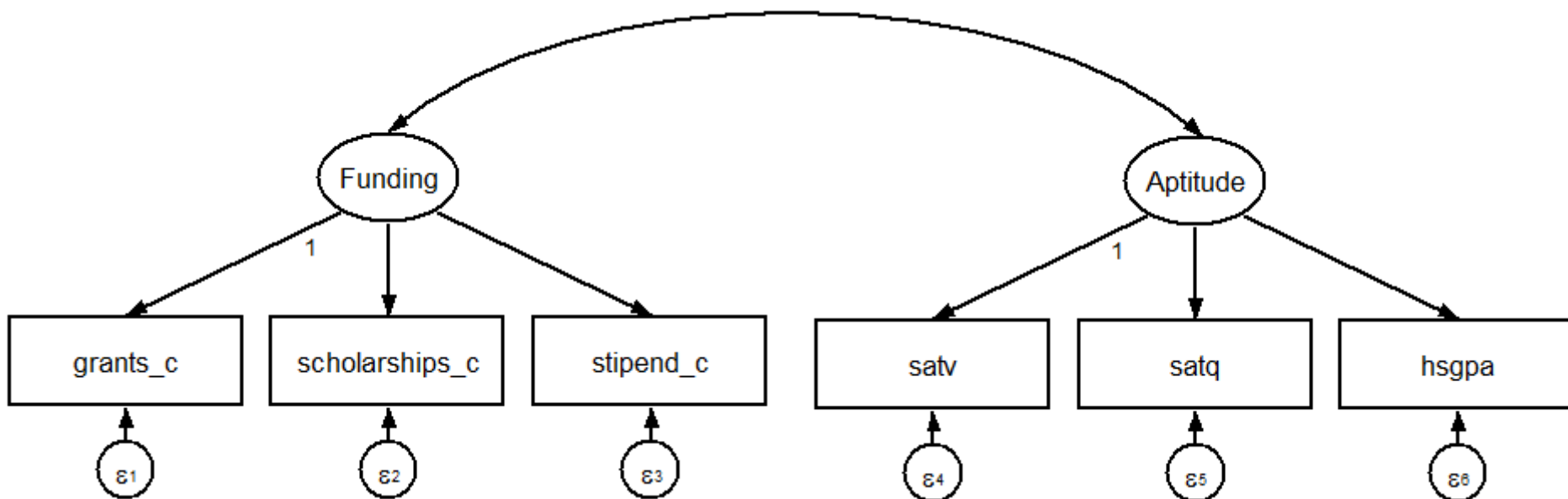
# Confirmatory Factory Analysis Path Diagram

```
sem (Funding -> grants_c scholarships_c stipend_c), latent(Funding)
```

	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Measurement						
grants_c <- Funding _cons	1 .0032446	(constrained) .008761	0.37	0.711	-.0139266	.0204158
scholars~c <- Funding _cons	1.249911 -.0097801	.0223002 .0087471	56.05 -1.12	0.000 0.264	1.206203 -.026924	1.293618 .0073639
stipend_c <- Funding _cons	.8651043 -.0099026	.0151705 .0089287	57.03 -1.11	0.000 0.267	.8353706 -.0274024	.8948379 .0075973
var(e.grants_c)	.5185624	.0097522			.4997964	.538033
var(e.scholar~c)	.2477532	.0118448			.2255924	.2720911
var(e.stipend_c)	.6767664	.0100357			.6573799	.6967245
var(Funding)	.4760264	.0128365			.4515206	.5018622

LR test of model vs. saturated: chi2(0) = 0.00, Prob > chi2 = .

# Confirmatory Factor Analysis Path Diagram



```
sem (Funding -> grants_c@1 scholarships_c stipend_c) ///  
    (Aptitude -> satv@1 satq hsgpa),           ///  
latent(Funding Aptitude)                       ///  
cov( Funding*Aptitude)
```

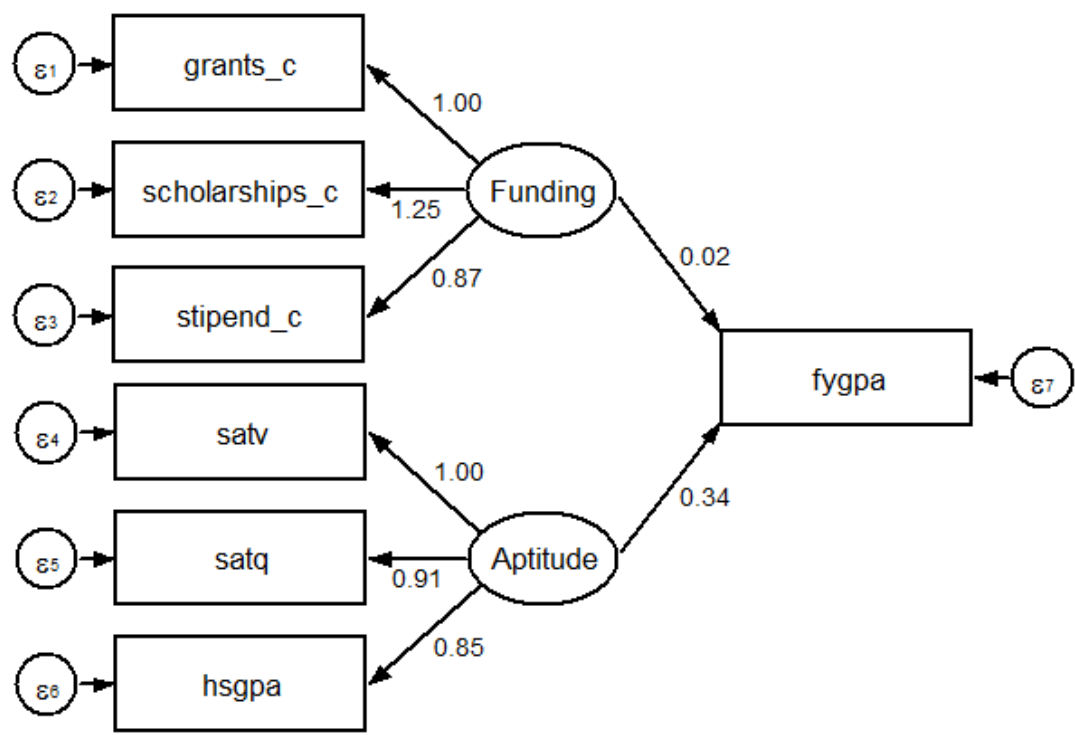
	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Measurement grants_c <- Funding _cons	1 .002366	(constrained) .0087934	0.27	0.788	-.0148688	.0196009
scholar~c <- Funding _cons	1.250233 -.009954	.0223636 .0087752	55.90 -1.13	0.000 0.257	1.206401 -.027153	1.294065 .007245
stipend_c <- Funding _cons	.8652328 -.0112386	.0152112 .0089581	56.88 -1.25	0.000 0.210	.8354195 -.0287962	.8950462 .006319
satv <- Aptitude _cons	1 1.170703	(constrained) .0093232	125.57	0.000	1.15243	1.188976
satq <- Aptitude _cons	.9835765 .8074195	.0161097 .0090827	61.06 88.90	0.000 0.000	.9520021 .7896178	1.015151 .8252212
hsgpa <- Aptitude _cons	.924639 .5970783	.0151946 .0089403	60.85 66.79	0.000 0.000	.894858 .5795557	.9544199 .614601
var(e.grants_c)	.5192783	.0097869			.5004463	.538819
var(e.scholar~c)	.2469635	.0118817			.2247401	.2713844
var(e.stipend_c)	.6766382	.0100666			.6571929	.6966588
var(e.satv)	.5220747	.0105282			.5018423	.5431228
var(e.satq)	.484536	.0100291			.4652728	.5045968
var(e.hsgpa)	.5186371	.0095846			.5001879	.5377669
var(Funding)	.4762764	.012883			.451684	.5022079
var(Aptitude)	.5970382	.0148564			.5686188	.626878
cov(Funding, Aptitude)	-.0048916	.0059336	-0.82	0.410	-.0165213	.0067381

LR test of model vs. saturated:  $\chi^2(8) = 5.16$ , Prob >  $\chi^2 = 0.7408$

# Continuous Outcome Models Using `sem`

- Example Data
- Means
- Correlation
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- **Structural Equation Models (SEM)**
- Multi-group SEM
- SEM For Complex Survey Data

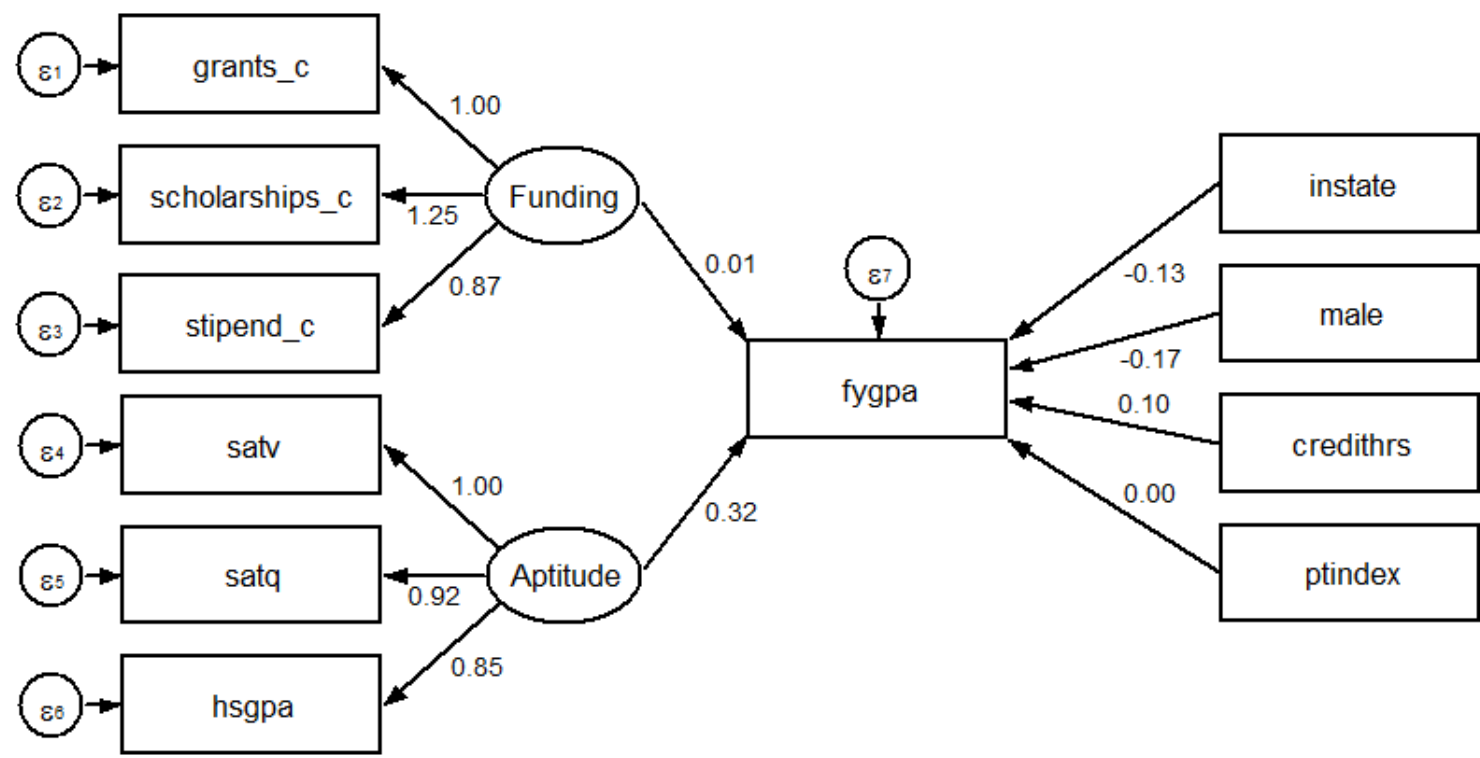
# Structural Equation Model Path Diagram



```

sem (Funding -> grants_c@1 scholarships_c stipend_c) ///
    (Aptitude -> satv@1 satq hsgpa)                ///
    (Funding Aptitude -> fygpa),                  ///
latent(Funding Aptitude)
  
```

# Structural Equation Model Path Diagram



```

sem (Funding -> grants_c@1 scholarships_c stipend_c)          ///
    (Aptitude -> satv@1 satq hsgpa)                        ///
    (Funding Aptitude -> fygpa)                            ///
    (instate male credithrs ptindex -> fygpa),            ///
    latent(Funding Aptitude)
  
```



# Structural Equation Models

Getting complex models to converge can sometimes be challenging. It may help to fit the full model in stages using the results of each simpler model as the starting values for more complex models:

```
sem (Funding -> grants_c@1 scholarships_c stipend_c)      ///
    (Aptitude -> satv@1 satq hsgpa)                    ///
    (Funding Aptitude -> fygpa),                       ///
    latent(Funding Aptitude)
```

```
matrix b = e(b)
```

```
sem (Funding -> grants_c@1 scholarships_c stipend_c)      ///
    (Aptitude -> satv@1 satq hsgpa)                    ///
    (Funding Aptitude -> fygpa)                       ///
    (instate male credithrs ptindex -> fygpa),        ///
    latent(Funding Aptitude)                          ///
    from(b)
```

# Structural Equation Models

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

Fit statistic	value	Description
Likelihood ratio		
chi2_ms(28)	411.457	model vs. saturated
p > chi2	0.000	
chi2_bs(49)	22294.001	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.033	Root mean squared error of approximation
90% CI, lower bound	0.030	
upper bound	0.035	
pclose	1.000	Probability RMSEA <= 0.05
Information criteria		
AIC	410228.932	Akaike's information criterion
BIC	410490.138	Bayesian information criterion
Baseline comparison		
CFI	0.983	Comparative fit index
TLI	0.970	Tucker-Lewis index
Size of residuals		
SRMR	0.013	standardized root mean squared residual
CD	0.961	Coefficient of determination

The goodness of fit statistics indicate that our models fits well

# Structural Equation Models

```
. estat residuals, format(%4.1f)
```

Residuals of observed variables

Mean residuals

	gran	scho	stip	satv	satq	hsgp	fygp	inst	male	cred	ptin
raw	-0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Covariance residuals

	gran	scho	stip	satv	satq	hsgp	fygp	inst	male	cred	ptin
grants_c	-0.0										
scholarshi~c	0.0	0.0									
stipend_c	-0.0	-0.0	0.0								
satv	-0.0	0.0	-0.0	0.0							
satq	0.0	0.0	-0.0	-0.0	0.0						
hsgpa	0.0	-0.0	-0.0	-0.0	0.0	0.0					
fygpa	-0.0	0.0	-0.0	0.1	-0.0	-0.1	0.0				
instate	-0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	0.0			
male	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.0		
credithrs	-0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	
ptindex	-0.2	0.1	-0.1	0.1	0.0	-0.1	0.0	0.0	0.0	0.0	0.0

The residuals are small or zero

```
. estat mindices
```

Modification indices

	MI	df	P>MI	EPC	Standard EPC
<b>Structural</b>					
fygpa <-					
satv	332.417	1	0.00	.2245471	.3265529
satq	32.381	1	0.00	-.0601799	-.0852606
hsgpa	197.553	1	0.00	-.1341139	-.1870284
<b>Measurement</b>					
scholarships_c <- ptindex	5.791	1	0.02	.0010236	.0185995
<b>satv &lt;-</b>					
satq	219.539	1	0.00	-.5028531	-.4898824
hsgpa	33.981	1	0.00	-.1606859	-.1540867
fygpa	363.321	1	0.00	.2438424	.167673
instate	5.401	1	0.02	-.0414326	-.0173591
male	7.807	1	0.01	-.0449951	-.0208948
credithrs	11.401	1	0.00	.0260854	.0252267
<b>satq &lt;-</b>					
satv	219.539	1	0.00	-.558623	-.5734137
hsgpa	364.385	1	0.00	.4616847	.454446
fygpa	34.087	1	0.00	-.0711863	-.0502458
<b>hsgpa &lt;-</b>					
satv	33.981	1	0.00	-.1936382	-.2019313
satq	364.384	1	0.00	.5008192	.5087967
fygpa	218.432	1	0.00	-.1762143	-.1263594
instate	10.294	1	0.00	.0550984	.0240733
male	6.309	1	0.01	.0389548	.0188646
credithrs	4.963	1	0.03	-.0165768	-.0167176
<b>cov(e.satv,e.satq)</b>					
cov(e.satv,e.hsgpa)	33.981	1	0.00	-.089045	-.1763945
cov(e.satv,e.fygpa)	332.417	1	0.00	.1032584	.23088
cov(e.satq,e.hsgpa)	364.384	1	0.00	.2558447	.4808539
cov(e.satq,e.fygpa)	32.381	1	0.00	-.0307431	-.0652183
cov(e.hsgpa,e.fygpa)	197.553	1	0.00	-.0743199	-.1513768

EPC = expected parameter change

# Continuous Outcome Models Using `sem`

- Example Data
- Means
- Correlation
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- **Multi-group SEM**
- SEM For Complex Survey Data

# Multigroup SEM

We can also fit models by group and test for invariance of parameters across groups.

```
sem (Funding -> grants_c@1 scholarships_c stipend_c)    ///
    (Aptitude -> satv@1 satq hsgpa)                    ///
    (Funding Aptitude -> fygpa)                        ///
    (instate male credithrs ptindex -> fygpa),        ///
    latent(Funding Aptitude)                          ///
    group(private)
```

```
estat ggof
```

```
estat ginvariant
```

# Continuous Outcome Models Using `sem`

- Example Data
- Means
- Correlation
- Linear Regression
- Multivariate Regression
- Path Analysis and Mediation
- Confirmatory Factor Analysis (CFA)
- Structural Equation Models (SEM)
- Multi-group SEM
- **SEM For Complex Survey Data**

# SEM For Complex Survey Data

- We can use **sem** to fit models for data that were collected using complex probability samples.
- For example, we might have collected our data by drawing a sample of universities and then colleges within universities.
- We can tell Stata about these features using **svy set** and our models will be estimated correctly.



# SEM For Complex Survey Data

The screenshot shows the 'svyset - Survey data settings' dialog box in STATA. The 'Main' tab is selected. The 'Number of stages' is set to 2. The 'Clear settings' button is visible. The settings for Stage 1 are: Sampling units: university, Strata: private, Finite pop. correction: univ\_fpc. The settings for Stage 2 are: Sampling units: college, Strata: |, Finite pop. correction: coll\_fpc. A note at the bottom states: 'Note: empty or "\_n" in "Sampling units" above indicates sampling of observations.' The dialog box has 'OK', 'Cancel', and 'Submit' buttons at the bottom.

svyset - Survey data settings

Main More Weights SE Poststratification

Number of stages: 2

	Sampling units	Strata	Finite pop. correction
Stage 1:	university	private	univ_fpc
Stage 2:	college		coll_fpc

Note: empty or "\_n" in "Sampling units" above indicates sampling of observations.



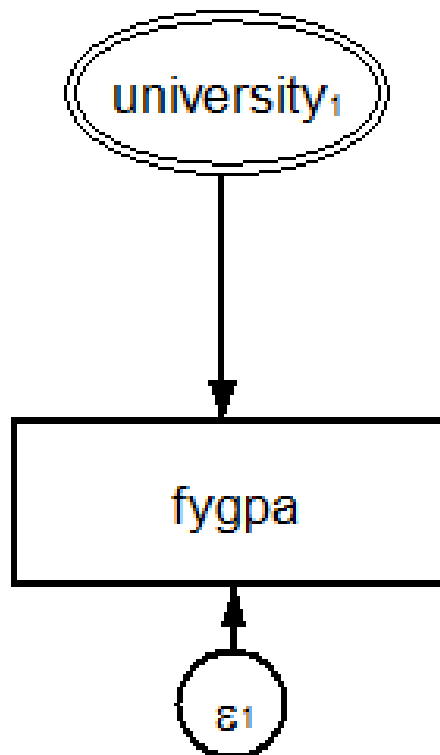
# Outline

- Introduction to Stata
- What is structural equation modeling?
- Structural equation modeling in Stata
- Continuous outcome models using **sem**
- **Multilevel generalized models using gsem**
- Demonstrations and Questions

# Multilevel Generalized Models Using `gsem`

- Multilevel models
- Multilevel CFA
- Logistic regression
- Generalized CFA
- Multilevel Generalized CFA
- Multilevel Generalized SEM

# Variance Component Model Path Diagram



# Variance Component Model Syntax

Syntax using **mixed**:

```
mixed fygpa || university:
```

Syntax using **gsem**:

```
gsem (M1[university] -> fygpa), latent(M1)
```

# Variance Component Model Results

Results using **mixed**:

fygpa	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	2.353906	.0903044	26.07	0.000	2.176912	2.530899

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
university: Identity var(_cons)	.1626113	.0515747	.0873332	.3027766
var(Residual)	.3127936	.0039015	.3052395	.3205348

Results using **gsem**:

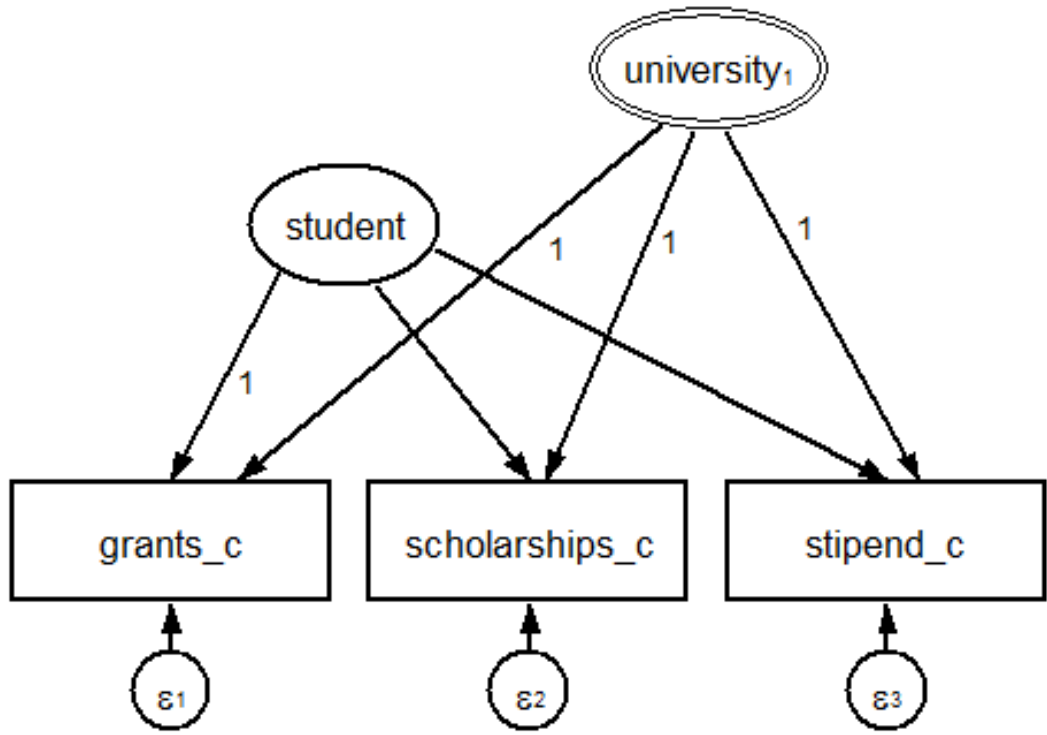
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
fygpa <- M1[university]	1 (constrained)					
_cons	2.353906	.0903044	26.07	0.000	2.176912	2.530899
var(M1[unive~y])	.1626113	.0515747			.0873331	.3027768
var(e.fygpa)	.3127936	.0039015			.3052395	.3205348

# Multilevel Generalized Models Using `gsem`

- Multilevel models
- **Multilevel CFA**
- Logistic regression
- Generalized CFA
- Multilevel Generalized CFA
- Multilevel Generalized SEM



# Multilevel CFA Path Diagram



# Multilevel CFA Results

```

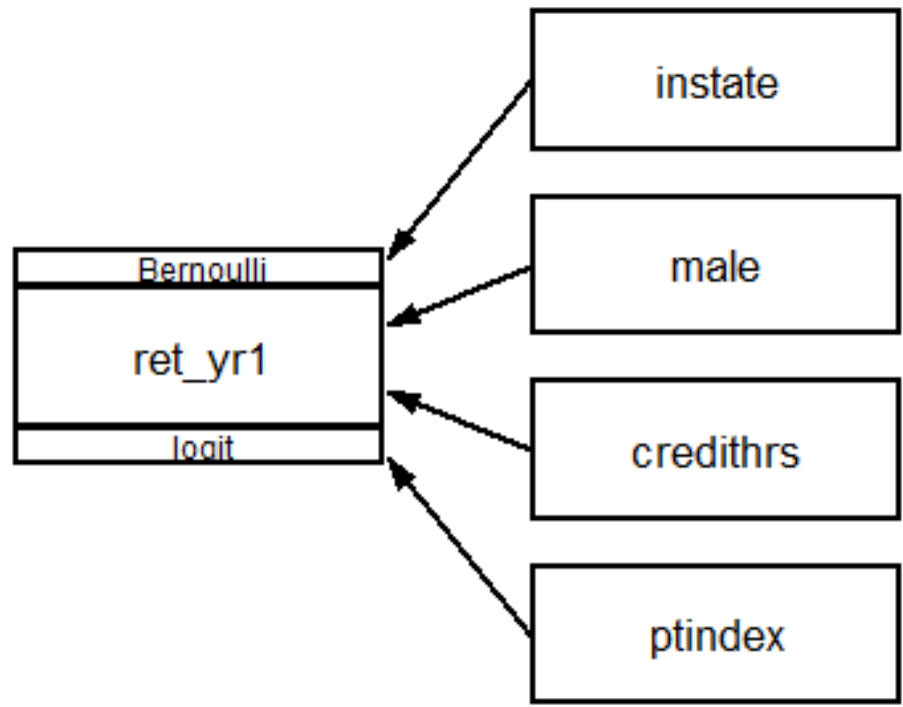
gsem (student -> grants_c@1 scholarships_c stipend_c)          ///
      (M1[university]@1 -> grants_c scholarships_c stipend_c),  ///
      covstruct(_lexogenous, diagonal) from(b) latent(student M1)  ///
      means(student@0 M1[university]@0) nocapslatent
  
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
grants_c <- M1[university]	1	(constrained)				
student _cons	-.0041953	.0087432	-0.48	0.631	-.0213317	.0129411
scholarshi~c <- M1[university]	1	(constrained)				
student _cons	1.175532 -.0064351	.0180882 .0087012	64.99 -0.74	0.000 0.460	1.14008 -.0234891	1.210985 .010619
stipend_c <- M1[university]	1	(constrained)				
student _cons	.821042 -.0086983	.0146714 .00877	55.96 -0.99	0.000 0.321	.7922866 -.0258873	.8497974 .0084907
var(M1[unive~y]) var(student)	7.99e-11 .4963372	. .0125068			. .4724198	. .5214656
var(e.grants_c) var(e.scholar~c) var(e.stipend_c)	.4950738 .2964019 .6626266	.0092272 .0089246 .0096384			.4773152 .2794162 .6440023	.5134932 .3144201 .6817894

# Multilevel Generalized Models Using `gsem`

- Multilevel models
- Multilevel CFA
- **Logistic regression**
- Generalized CFA
- Multilevel Generalized CFA
- Multilevel Generalized SEM

# Logistic Regression Path Diagram



# Logistic Regression Syntax

Syntax using `logit` or `logistic`:

```
logit ret_yr1 instate male credithrs ptindex
```

```
logistic ret_yr1 instate male credithrs ptindex
```

Syntax using `gsem`:

```
gsem ret_yr1 <- instate male credithrs ptindex, ///  
    family(bernoulli) link(logit)
```

```
gsem ret_yr1 <- instate male credithrs ptindex, logit
```

```
estat eform
```

# Logistic Regression Results

Results using `logistic`:

ret_yr1	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
instate	1.898841	.1134222	10.74	0.000	1.689057	2.134681
male	1.093629	.0643822	1.52	0.128	.9744497	1.227384
credithrs	1.376276	.0400066	10.99	0.000	1.300056	1.456964
ptindex	.9984765	.0016018	-0.95	0.342	.9953419	1.001621
_cons	.0394908	.0184686	-6.91	0.000	.0157912	.0987588

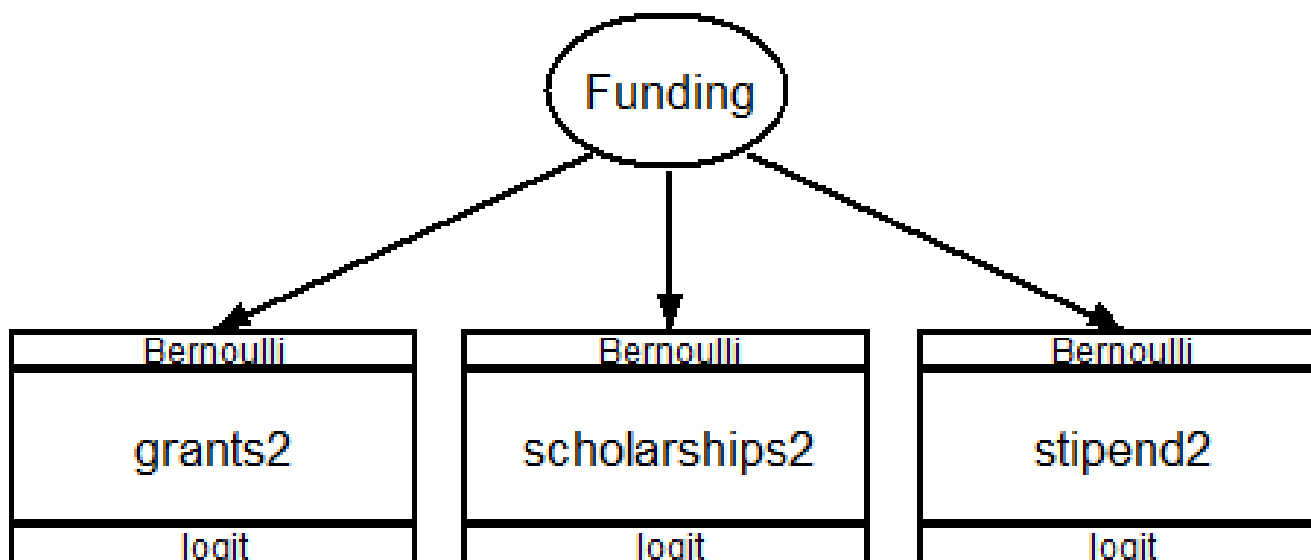
Results using `gsem` and `estat eform`:

ret_yr1	exp(b)	Std. Err.	z	P> z	[95% Conf. Interval]	
instate	1.898841	.1134222	10.74	0.000	1.689057	2.134681
male	1.093629	.0643822	1.52	0.128	.9744497	1.227384
credithrs	1.376276	.0400066	10.99	0.000	1.300056	1.456964
ptindex	.9984765	.0016018	-0.95	0.342	.9953419	1.001621
_cons	.0394908	.0184686	-6.91	0.000	.0157912	.0987588

# Multilevel Generalized Models Using `gsem`

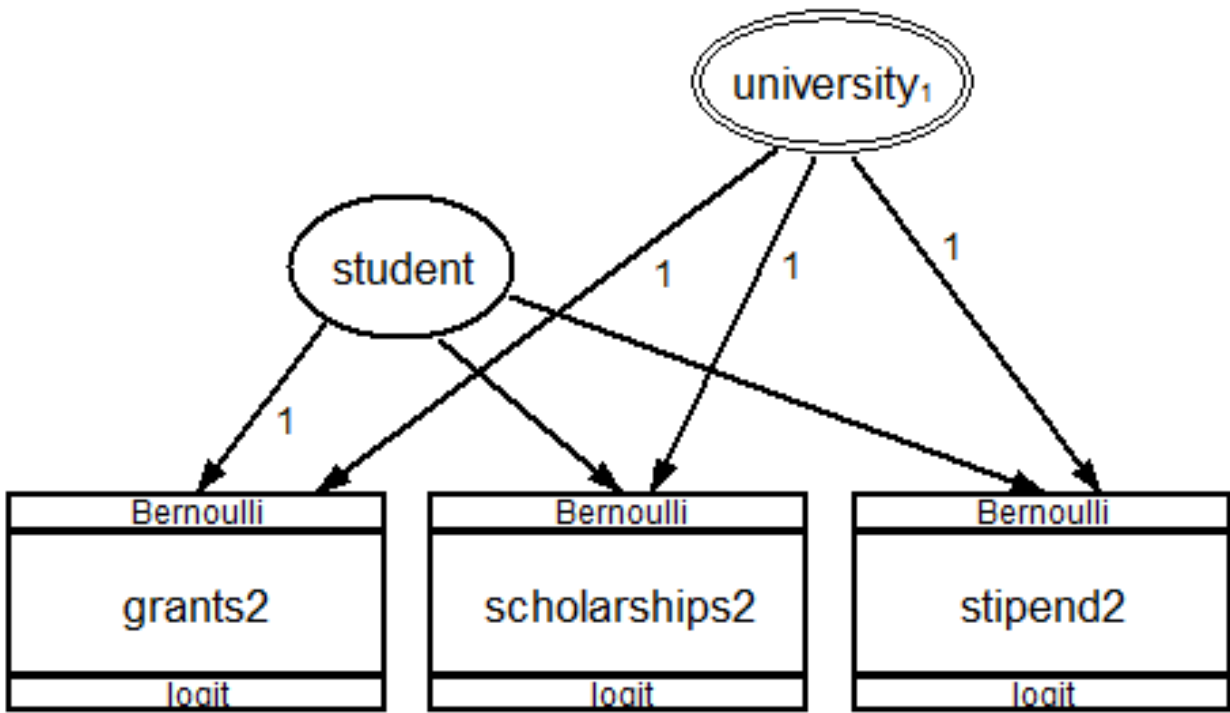
- Multilevel models
- Multilevel CFA
- Logistic regression
- **Generalized CFA**
- **Multilevel Generalized CFA**
- **Multilevel Generalized SEM**

# Generalized CFA Path Diagram

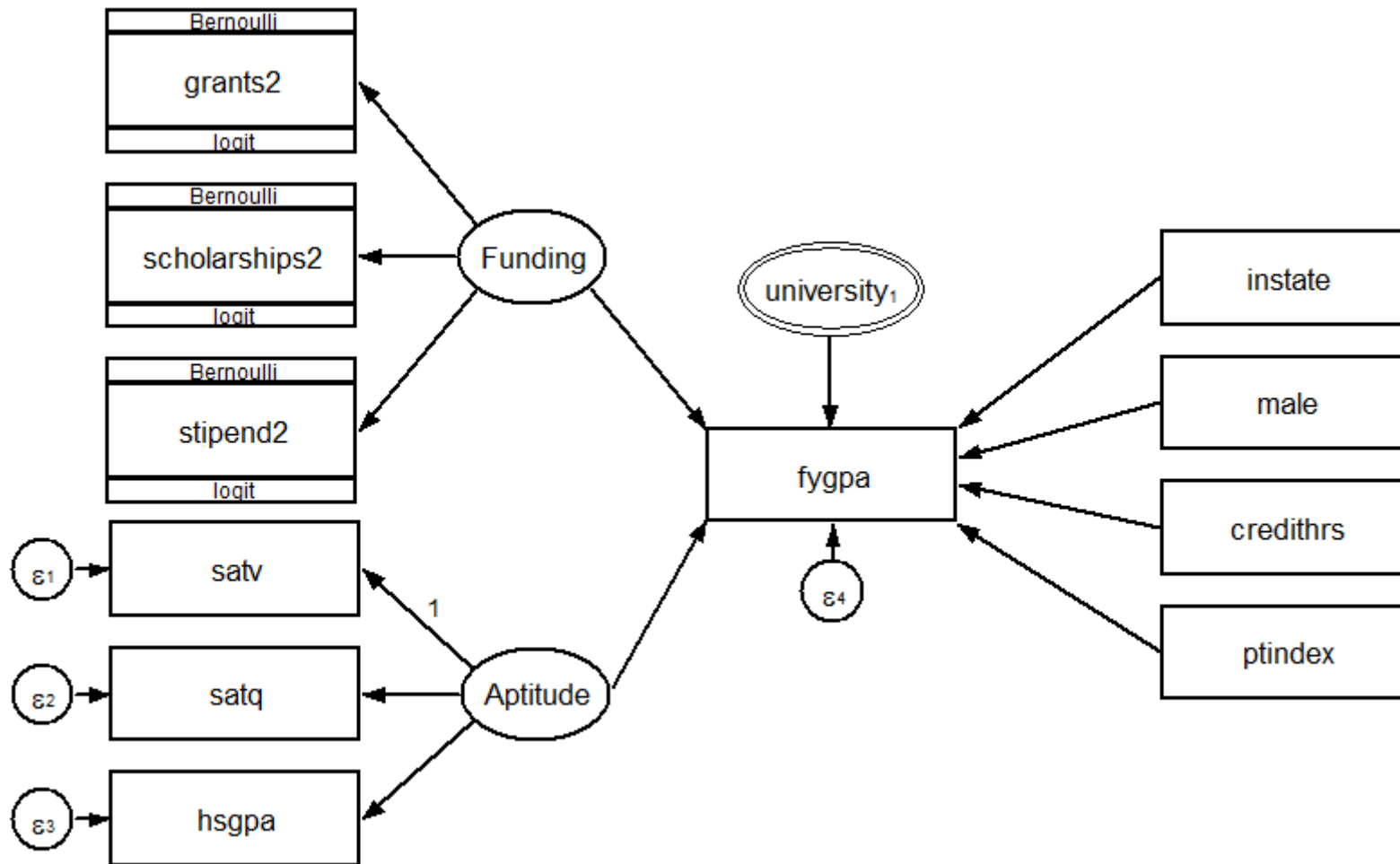




# Multilevel Generalized CFA Path Diagram



# Multilevel Generalized SEM Path Diagram



# Multilevel Generalized Models Using `gsem`

- Multilevel models
- Multilevel CFA
- Logistic regression
- Generalized CFA
- Multilevel Generalized CFA
- Multilevel Generalized SEM

# Outline

- Introduction to Stata
- What is structural equation modeling?
- Structural equation modeling in Stata
- Continuous outcome models using **sem**
- Multilevel generalized models using **gsem**
- **Demonstrations and Questions**

# Acknowledgements

- Kristin MacDonald
  - Senior Statistician, StataCorp
- Rose Medeiros
  - Senior Statistician, StataCorp
- Bryce Mason
  - Assistant Vice Chancellor, UC Riverside

# References and Further Reading

1. Stata 13 Structural Equation Modeling Reference Manual:  
[www.stata.com/manuals13/sem.pdf](http://www.stata.com/manuals13/sem.pdf)
2. Acock, A.C. (2013) *Discovering Structural Equation Modeling Using Stata*, Revised Edition . College Station, TX: Stata Press.
3. Bollen, K.A. (1989) *Structural Equations With Latent Variables*. New York: Wiley
4. Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
5. Kline, R.B. (2010). *Principles and Practice of Structural Equation Modeling* , 3<sup>rd</sup> Ed. New York: Guilford Press
6. Matsueda, R.L. (2012). Key Advances in the History of Structural Equation Modeling. *Handbook of Structural Equation Modeling*. 2012. Edited by R. Hoyle. New York, NY: Guilford Press
7. Rabe-Hesketh, S., and A. Skrondal. (2012) *Multilevel and Longitudinal Modeling Using Stata*. 3<sup>rd</sup> ed. College Station, TX: Stata Press.

# Demonstrations And Questions