

Introduction to SEM in Stata

Christopher F Baum

ECON 8823: Applied Econometrics

Boston College, Spring 2016

Introduction

We now present an introduction to Stata's `sem` command, which implements structural equation modeling. As `sem` has a very broad set of capabilities, we can only discuss a limited subset of its features and give some illustrations of its use in the time available. We also will not discuss the graphical interface to `sem`, the SEM Builder, but you are welcome to explore its capabilities for specifying the model graphically rather than in the command language.

Structural equation modeling allows us to combine *measurement models*, which involve the relationships between observed measurements and *latent*, or unobserved variables, with *path analysis* models that relate variables to their causal factors.

As an applied econometrician, rather than a psychologist or sociologist, I found the terminology used in SEM to be quite foreign to what we usually consider in economic modeling. However, digging deeper, I recognize the similarities.

For instance, we motivate the use of the binomial probit model in studying behavior: for instance, whether or not someone makes a purchase. We argue that the individual is calculating the expected net benefit of her action, which we cannot observe, but we observe the outcome of their decision process.

If the expected net benefit is positive, we observe a 1; if it is negative or zero, we observe a 0. In this case, expected net benefit is the underlying *latent variable* driving the decision process, and we can only observe its presumed sign, not its magnitude. So the concepts underlying a *measurement model* are perhaps not as foreign as some might think.

What is a *path analysis* model? As it turns out, another terminology for the sort of model used every day in applied econometrics, usually via some sort of regression techniques. The model is comprised of one or more equations (which, confusingly, are called *structural equations*) linking outcome variables (dependent variables, or *endogenous* variables) with causal factors (independent variables, or *exogenous* variables). In this context, all variables are presumed to be observable.

Structural equation models (SEM), then, combine these two types of model and allow for both latent variables, driven by observables, and relationships among observables. In that context, they often involve several equations, going beyond the common single-equation modeling strategy employed in much of applied econometrics. But as StataCorp's developers have pointed out, the SEM framework encompasses most of the techniques in common use in applied econometrics, while providing a number of useful extensions to several common methodologies.

The scope of SEM is very well put by Stata's introduction to SEM:
"Structural equation modeling is not just an estimation method for a particular model in the way that Stata's `regress` and `probit` commands are, or even in the way that `stcox` and `mixed` are. Structural equation modeling is a way of thinking, a way of writing, and a way of estimating." ([SEM] 2).

One other tribal distinction in the application of SEM is a preference among some tribes for working with these models' graphical representations. Stata's SEM Builder provides full support for that strategy, allowing you to both 'draw' the model and express the interrelationships in the diagram and then estimate the model as illustrated. The results of estimation are then displayed on the drawing, which can be produced in publication-quality form.

Given my unfamiliarity with other SEM software, I cannot attest to the ease of use or quality of output provided by SEM Builder relative to that of competing products. I will not focus on the SEM Builder approach in these talks, largely due to my own unfamiliarity with it and that mode of working (I don't use menus, dialogs, etc. in working with Stata, either). But for those who like to draw their models, I suggest that Stata's facility for doing so is well worth learning.

A classic example of SEM modeling

To motivate the full SEM framework, we present a classic example of structural equation modeling, as discussed by Acock in *Discovering Structural Equation Modeling using Stata*.¹ This is a model developed by Wheaton et al. (*Sociological Methodology* 1977) to analyze the concept of individuals' alienation.

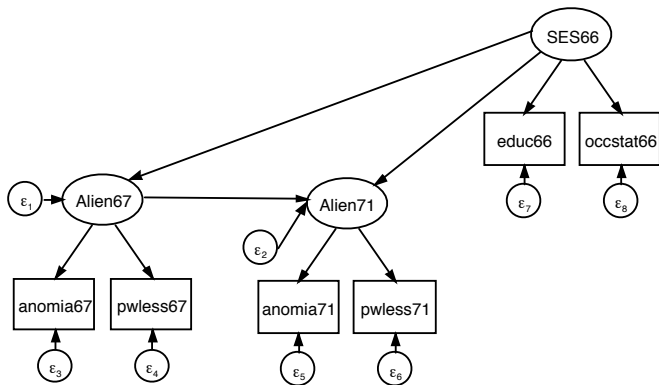
¹A revised edition of this book was published by Stata Press in 2013.

Two latent variables are the object of investigation: alienation in 1967 and alienation in 1971. A third latent variable, socioeconomic status (SES) in 1966, also plays a role in the model. The underlying data contain information on two measures thought to reflect socioeconomic status: level of education and occupational status, both measured in 1996.

Survey responses for two factors, anomia² and powerlessness, were measured in 1967 and again in 1971. Those are taken as indicators of alienation. Additionally, as the key research question regards the stability of alienation, alienation in the earlier year (1967) is thought to have a causal relationship with alienation in the later year (1971).

²A difficulty in remembering the meaning of words.

To illustrate this model graphically:



Note that capitalized variable names refer to latent variables, while lower case names are observed variables. There are three measurement equations, for Alien67, Alien71, and SES66. The observed measures should reflect their respective latent variables. Hence, the arrows point to the observed measures. Alien67 is taken as related to SES66, and Alien71 is taken as depending on both Alien67 and SES66.

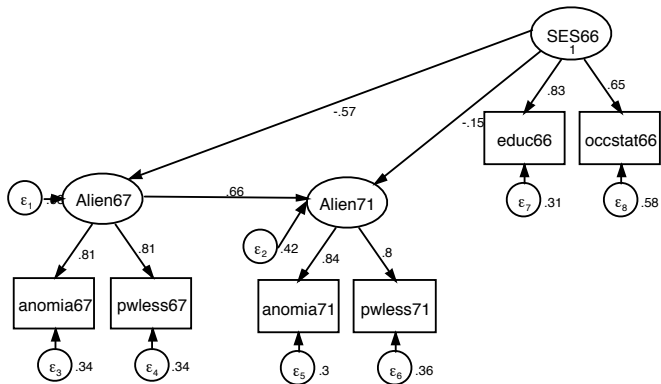
In Stata's command language, this model can be specified as:

```
use http://www.stata-press.com/data/r13/sem_sm2.dta, clear
sem ///
(Alien67 -> anomia67 pwless67)   /// measure Alien67
(Alien71 -> anomia71 pwless71)   /// measure Alien71
(SES66 -> educ66 occstat66)     /// measurement piece
(Alien67 <- SES66)              /// structural piece
  (Alien71 <- Alien67 SES66),   /// structural piece
standardized                    // Options
```

SEM can be used where we only have the summary statistics of the data: means and their covariance (or correlation) matrix. In this model, we have 6 observed variables, or indicators. Their variance-covariance matrix (VCE) thus contains $6(6+1) / 2 = 21$ elements: 6 variances and 15 covariances. The degrees of freedom of our estimated model will reflect the number of parameters to be estimated (variances of the latent factors, variances of the error terms, and path coefficients). In this context, with several parameters set to 1.0, we have 15 parameters to be estimated, and thus 6 degrees of freedom.

Stata will consider that the indicators in the measurement model, as well as the two latent alienation variables, are endogenous in the estimation, while SES66 is considered as an exogenous latent variable, affecting each alienation variable but not being affected by those variables.

Whether we estimate the model within SEM Builder or via the command language, we will get the same results:




```

.
. sem
> (Alien67 -> anomia67 pwless67) /// measure Alien67
> (Alien71 -> anomia71 pwless71) /// measure Alien71
> (SES66 -> educ66 occstat66) /// measurement piece
> (Alien67 <- SES66) /// structural piece
> (Alien71 <- Alien67 SES66), /// structural piece
> standardized nolog // Options

Endogenous variables
Measurement:  anomia67 pwless67 anomia71 pwless71 educ66 occstat66
Latent:      Alien67 Alien71

Exogenous variables
Latent:      SES66

Structural equation model          Number of obs      =          932
Estimation method = ml
Log likelihood      = -15246.469

( 1) [anomia67]Alien67 = 1
( 2) [anomia71]Alien71 = 1
( 3) [educ66]SES66 = 1

```

Standardized	Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural Alien67 <- SES66	-.5668218	.0344036	-16.48	0.000	-.6342517	-.4993919
Alien71 <- Alien67	.6630088	.0396724	16.71	0.000	.5852523	.7407654
SES66	-.151492	.0458162	-3.31	0.001	-.24129	-.061694
Measurement anomia67 <- Alien67	.812882	.0194328	41.83	0.000	.7747943	.8509697
_cons	3.95852	.097363	40.66	0.000	3.767692	4.149347
pwless67 <- Alien67	.811926	.0194466	41.75	0.000	.7738113	.8500406
_cons	4.796692	.1158294	41.41	0.000	4.56967	5.023713
anomia71 <- Alien71	.8395125	.0193263	43.44	0.000	.8016337	.8773913
_cons	3.993669	.09813	40.70	0.000	3.801338	4.186

<hr/>						
pwless71 <-						
Alien71	.798082	.0198613	40.18	0.000	.7591546	.8370095
_cons	4.717723	.1140761	41.36	0.000	4.494137	4.941308
<hr/>						
educ66 <-						
SES66	.8326718	.031738	26.24	0.000	.7704664	.8948772
_cons	3.518017	.0878219	40.06	0.000	3.345889	3.690145
<hr/>						
occstat66 <-						
SES66	.6485148	.0301669	21.50	0.000	.5893887	.707641
_cons	1.767678	.0524337	33.71	0.000	1.66491	1.870446
<hr/>						
var(e.anomia67)	.3392229	.0315932			.2826241	.4071562
var(e.pwless67)	.3407762	.0315784			.2841788	.4086457
var(e.anomia71)	.2952187	.0324493			.2380034	.3661885
var(e.pwless71)	.3630651	.0317019			.3059565	.4308333
var(e.educ66)	.3066577	.0528548			.2187474	.4298974
var(e.occstat66)	.5794285	.0391274			.5075984	.6614233
var(e.Alien67)	.6787131	.0390015			.6064191	.7596255
var(e.Alien71)	.4236057	.0345717			.360988	.4970851
var(SES66)	1	.			.	.
<hr/>						

LR test of model vs. saturated: chi2(6) = 71.62, Prob > chi2 = 0.0000

As we would expect, the effect of higher SES66 on alienation in each year is negative and significant, with a stronger impact on the near term (1967) value than on the longer-term value (1971). The link between alienation in the two years is also positive and significant, suggesting the presence of stability in individuals' attitudes.

We may next examine the goodness-of-fit statistics to evaluate how much of the variance of each endogenous variable is being explained by the model.

```
. estat eggof          // R-squares
Equation-level goodness of fit
```

depvars	fitted	Variance predicted	residual	R-squared	mc	mc2
observed						
anomia67	11.8209	7.810982	4.009921	.6607771	.812882	.6607771
pwless67	9.353552	6.166084	3.187468	.6592238	.811926	.6592238
anomia71	12.51815	8.822558	3.695593	.7047813	.8395125	.7047813
pwless71	9.974882	6.35335	3.621531	.6369349	.798082	.6369349
educ66	9.599689	6.65587	2.943819	.6933423	.8326718	.6933423
occstat66	449.8053	189.1753	260.63	.4205715	.6485148	.4205715
latent						
Alien67	7.810982	2.509567	5.301416	.3212869	.5668218	.3212869
Alien71	8.822558	5.085272	3.737286	.5763943	.7592064	.5763943
overall				.7784845		

mc = correlation between depvar and its prediction

mc2 = mc² is the Bentler-Raykov squared multiple correlation coefficient

The model has explained 32.1% of the variance in the latent variable `Alien67` and 57.6% of the variance in the latent variable `Alien71`. The significant value of the coefficient linking the two measures suggests that there is substantial stability over the years. That estimate may be larger than that of earlier studies because the indicator variables' measurement error is being taken into account.

Although these results are promising, the Chi-squared value from the estimation suggests that we are not doing a very good job of fitting the original covariance matrix. Unlike regression or logistic regression, where the summary statistic should reject its null to indicate validity of the model, the Chi-squared statistic reported in SEM output, a likelihood-ratio (LR) statistic comparing the model to the saturated model, will not reject its null if the model is adequate.

Improving the model

We consider how the model might be improved. Guidance in this task can be provided by *modification indices* (`estat mindices`), which measure how much the Chi-squared statistic would be altered by modifying the specification. To make the model more complex, we must have sufficient degrees of freedom to estimate additional parameters. If you recall, there are 6 residual degrees of freedom in the current specification.

The first thing to consider is allowing the error terms of `anomia67` and `anomia71` to be correlated, as well as the error terms of `pwless67` and `pwless71`. By default, those error terms are assumed to have zero correlation. A rationale for these correlations might be the presence of additional, unobserved factors that influence the indicators, but are not themselves measurable.

We add these correlations to the model, referring to them as `cov(e.indic1*e.indic2)`, where the `e.` prefix stands for error.

```

. * adding correlated error terms
. sem
>      (anomia67 pwless67 <- Alien67)  /// measure Alien67
>      (anomia71 pwless71 <- Alien71)  /// measure Alien71
>      (SES66 -> educ66 occstat66)     /// measurement piece
>      (Alien67 <- SES66)               /// structural piece
>      (Alien71 <- Alien67 SES66),     /// structural piece
>      cov(e.anomia67*e.anomia71)      /// correlated error
>      cov(e.pwless67*e.pwless71)      /// correlated error
>      method(ml) standardized nolog   // options

```

Endogenous variables

Measurement: anomia67 pwless67 anomia71 pwless71 educ66 occstat66

Latent: Alien67 Alien71

Exogenous variables

Latent: SES66

Structural equation model

Number of obs = 932

Estimation method = ml

Log likelihood = -15213.046

(1) [anomia67]Alien67 = 1

(2) [anomia71]Alien71 = 1

(3) [educ66]SES66 = 1

Standardized	Coef.	OIM Std. Err.	z	P> z	[95% Conf. In	
Structural Alien67 <- SES66	-.5631417	.0347138	-16.22	0.000	-.6311794	-
Alien71 <- Alien67	.5670411	.0409739	13.84	0.000	.4867337	.
SES66	-.2076461	.0452784	-4.59	0.000	-.2963901	-.
Measurement anomia67 <- Alien67	.7745404	.0253584	30.54	0.000	.7248389	.
_cons	3.958737	.0973322	40.67	0.000	3.767969	4
pwless67 <- Alien67	.8520275	.0259381	32.85	0.000	.8011898	.
_cons	4.796617	.1158258	41.41	0.000	4.569603	5
anomia71 <- Alien71	.8055306	.0260403	30.93	0.000	.7544926	.
_cons	3.99335	.0980611	40.72	0.000	3.801154	4

pwless71 <-							
	Alien71	.8318689	.0267765	31.07	0.000	.7793879	.
	_cons	4.717814	.1140716	41.36	0.000	4.494238	4
educ66 <-							
	SES66	.8413924	.0320905	26.22	0.000	.7784962	.
	_cons	3.518017	.0878219	40.06	0.000	3.345889	3
occstat66 <-							
	SES66	.6417933	.0302822	21.19	0.000	.5824413	.
	_cons	1.767678	.0524337	33.71	0.000	1.66491	1
var(e.anomia67)		.4000872	.0392821			.3300505	.
var(e.pwless67)		.2740492	.0441999			.1997758	.
var(e.anomia71)		.3511205	.0419524			.2778134	.
var(e.pwless71)		.3079941	.0445491			.2319649	.
var(e.educ66)		.2920589	.0540014			.2032751	.
var(e.occstat66)		.5881014	.0388698			.516646	.
var(e.Alien67)		.6828714	.0390975			.6103848	.
var(e.Alien71)		.5027345	.0333311			.4414732	.
var(SES66)		1	.			.	.
cov(e.anomia67,e.anomia71)		.3557506	.0472739	7.53	0.000	.2630954	.
cov(e.pwless67,e.pwless71)		.1211569	.0819699	1.48	0.139	-.0395011	.

LR test of model vs. saturated: $\chi^2(4) = 4.78$, Prob > $\chi^2 = 0.3111$

In comparison to our earlier estimates, the effect of `SES66` on `Alien71` has increased, while the effect of `Alien67` on `Alien71` has decreased from 0.66 to 0.57, while being estimated very precisely. The covariance we have estimated between anomia error terms is positive and significant, while that for powerlessness is positive but not different from zero.

Most importantly, the model now fits adequately, with the p-value of the Chi-squared statistic rising to 0.31. By using two additional degrees of freedom, the model now more faithfully represents the relationships it encompasses.

We may also examine the goodness-of-fit statistics from this version of the model:

```
. estat eqgof
```

```
Equation-level goodness of fit
```

depvars	fitted	Variance predicted	residual	R-squared	mc	mc2
observed						
anomia67	11.81961	7.090733	4.728874	.5999128	.7745404	.5999128
pwless67	9.353843	6.79043	2.563413	.7259508	.8520275	.7259508
anomia71	12.52015	8.124068	4.396081	.6488795	.8055306	.6488795
pwless71	9.974493	6.902408	3.072085	.6920059	.8318689	.6920059
educ66	9.599688	6.796014	2.803674	.7079411	.8413924	.7079411
occstat66	449.8052	185.2742	264.5311	.4118986	.6417933	.4118986
latent						
Alien67	7.090733	2.248674	4.842059	.3171286	.5631417	.3171286
Alien71	8.124068	4.039819	4.084249	.4972655	.7051706	.4972655
overall				.7860745		

mc = correlation between depvar and its prediction

mc2 = mc² is the Bentler-Raykov squared multiple correlation coefficient

Im We now have lower R-squared terms for the two latent variables, as we are taking other factors into account in allowing for the error covariances to be nonzero.

Given the model's specification, `SES66` has both a direct effect on `Alien71` and an indirect effect, working through `Alien67`. We may request the indirect effects:

Indirect effects

	Coef.	OIM Std. Err.	z	P> z	Std. Coef.
Measurement					
anomia67 <-					
Alien67	0	(no path)			0
SES66	-.5752228	.057961	-9.92	0.000	-.436176
pwless67 <-					
Alien67	0	(no path)			0
SES66	-.5629103	.0507614	-11.09	0.000	-.4798122
anomia71 <-					
Alien67	.606954	.0512305	11.85	0.000	.456769
Alien71	0	(no path)			0
SES66	-.5761639	.059618	-9.66	0.000	-.424491
pwless71 <-					
Alien67	.5594603	.0472218	11.85	0.000	.4717039
Alien71	0	(no path)			0
SES66	-.5310796	.0516934	-10.27	0.000	-.4383705
educ66 <-					
SES66	0	(no path)			0
occstat66 <-					
SES66	0	(no path)			0

Structural						
Alien67 <-						
SES66	0	(no path)				0
Alien71 <-						
Alien67	0	(no path)				0
SES66	-.3491338	.0412546	-8.46	0.000		-.3193245

Here we see the estimate of that key indirect effect, SES66 → Alien71, as being negative and clearly significant.

The syntax of Stata's sem command

In Stata, you describe a SEM as a set of *paths*. Optionally, you may specify arguments to the `covariance()`, `variance()` and `means()` options.

The `covariance()` option is used to specify that a particular covariance path of the model is to be estimated, rather than being assumed 0; or that a nonzero covariance path is to be constrained to be 0, or some other fixed value. You may also constrain two or more covariance paths to be equal. The same features apply to the `variance()` and `means()` options.

In the path notation,

- 1 Latent variables are indicated by a name in which at least the first letter is capitalized.
- 2 Observed variables are indicated by a name in which at least the first letter is lowercased. Observed variables must correspond to variable names in the dataset.
- 3 Error variables, while mathematically a special case of latent variables, are considered in a class by themselves. For sem, every endogenous variable (whether observed or latent) automatically has an error variable associated with it. The error variable associated with endogenous variable *name* is *e.name*.
- 4 Paths between variables are written as $(name1 \leftarrow name2)$ or, alternatively, $(name2 \rightarrow name1)$.

- 5 Paths between the same variables can be combined: The paths $(name1 \leftarrow name2)$ $(name1 \leftarrow name3)$ can be combined as $(name1 \leftarrow name2\ name3)$
- 6 The paths $(name1 \leftarrow name2\ name3)$ $(name4 \leftarrow name2\ name3)$ may be written as $(name1\ name4 \leftarrow name2\ name3)$
- 7 Variances and covariances (curved paths) between variables are indicated by options. Variances are indicated by $\dots, \dots\ var(name1)$, while covariances are indicated by $\dots, \dots\ cov(name1*name2)$. Variances may be combined, covariances may be combined, and variances and covariances may be combined.
- 8 All variables except endogenous variables are assumed to have a variance; it is only necessary to code the $var()$ option if you wish to place a constraint on the variance or specify an initial value.

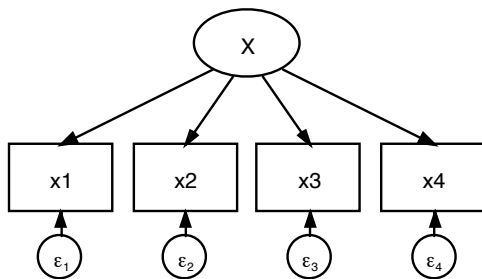
- 9 Endogenous variables have a variance, but that is the variance implied by the model. If `name` is an endogenous variable, then `var(name)` is invalid. The error variance of the endogenous variable is `var(e.name)`.
- 10 Variables mostly default to being correlated. All exogenous variables are assumed to be correlated with each other, whether observed or latent. Endogenous variables are never directly correlated, although their associated error variables can be. All error variables are assumed to be uncorrelated with each other. You can override these defaults on a variable-by-variable basis with the `cov()` option.

- 11 Variables mostly default to having nonzero means. All observed exogenous variables are assumed to have nonzero means. Latent exogenous variables are assumed to have mean 0. Endogenous variables have no separate mean. Their means are those implied by the model. Error variables have mean 0 and this cannot be modified. To constrain the mean to a fixed value, such as 57, code
..., ... means(name@57)
- 12 Fixed-value constraints may be specified for a path, variance, covariance, or mean by using @ (the “at” symbol). For example,
(name1 <- name2@1)
(name1 <- name2@1 name3@1)
..., ... var(name@100)

- 13 Symbolic constraints may be specified for a path, variance, covariance, or mean by using @. For example, $(name1 <- name2@c1)$ $(name3 <- name4@c1)$
Symbolic names are just names from 1 to 32 characters in length. Symbolic constraints constrain equality. For simplicity, all constraints below will have names $c1$, $c2$, ...
- 14 Linear combinations of symbolic constraints may be specified for a path, variance, covariance, or mean by using @. For example, $(name1 <- name2@c1)$ $(name3 <- name4@(2*c1))$
- 15 All equations in the model are assumed to have an intercept (to include observed exogenous variable *cons*) unless the *noconstant* option is specified.

Models supported by SEM

We now consider a number of models that are supported by the SEM methodology. The first is the *single-factor measurement model*, in which we consider several observed variables as influencing a single latent factor. This can be graphically represented:



In this model, we have four observed variables, each of which is presumed measured with error: hence the ϵ terms attached to each. They are presumed to relate to a single latent factor. Notice the notation, with capital letters denoting latent variables, and lowercase variable names for observed variables.

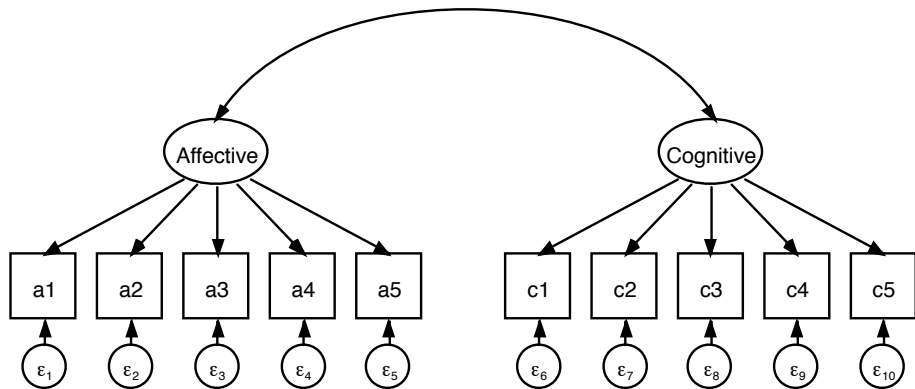
This may be estimated with `sem` as:

```
sem (x1 x2 x3 x4 <- X)
```

This is a pure measurement model, with no structural component.

The one-factor measurement model

We may extend this to the two-factor measurement model, where we have two latent factors, each related to a set of observed variables. We presume that the latent factors are correlated with one another, represented by the curved path in the diagram.



This may be estimated with `sem` as:

```
sem (Affective -> a1 a2 a3 a4 a5) (Cognitive -> c1 c2 c3 c4 c5)
```

This is a pure measurement model, with no structural component.

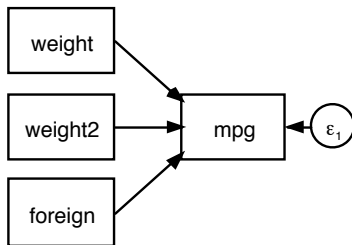
Linear regression

Linear regression is subsumed in the SEM framework as a pure structural model, with no measurement component nor latent variables:

```
regress mpg weight c.weight#c.weight foreign
```

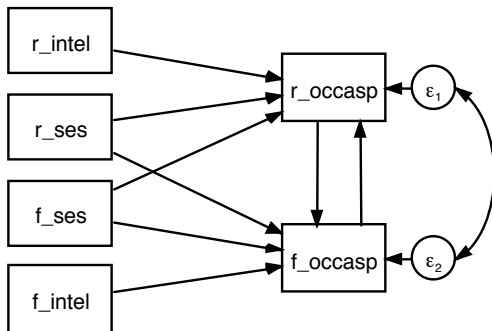
which may be related to the graphical representation. By default, `sem` will produce standardized coefficients, equivalent to those generated by `regress, beta`.

Here we have three exogenous variables and a single, continuous outcome variable, with a Gaussian error term.



Nonrecursive structural model

As a second example of a pure structural model, with all variables observed, we have a model from Duncan et al. (1968) which relates occupational aspirations of a respondent and his friend to several observed variables, including intelligence and socioeconomic status (SES). The SES measures of both individuals are hypothesized to affect each person's occupational aspirations. The occupational aspirations variables are assumed to be interrelated, with their error terms correlated.



This model may be fit with the `sem` command as:

```
sem (r_occasp <- f_occasp r_intel r_ses f_ses) ///  
    (f_occasp <- r_occasp f_intel f_ses r_ses), ///  
    cov(e.r_occasp*e.f_occasp) standardized
```

where the `r_` prefix stands for respondent and the `f_` prefix stands for friend.

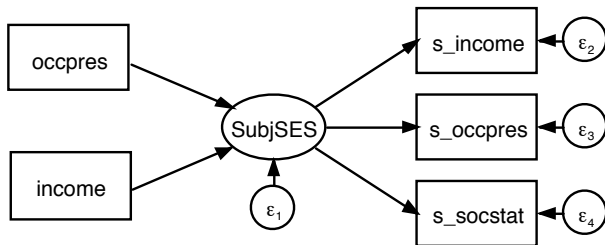
We can test whether coefficients in this model are equal to one another using the standard `test` command. Having determined that linear constraints are appropriate, we can reestimate the model:

```
sem (r_occasp <- f_occasp@b1 r_intel@b2 r_ses@b3 f_ses@b4) ///  
(f_occasp <- r_occasp@b1 f_intel@b2 f_ses@b3 r_ses@b4), ///  
cov(e.r_occasp*e.f_occasp)
```

where the symbolic names `b1`, `b2`, `b3`, `b4` indicate that single parameters are to be estimated for each of those names, and applied to the model. This will conserve degrees of freedom and increase the efficiency of estimation.

MIMIC model

We illustrate the MIMIC (multiple indicators, multiple causes) model. In this framework, using data from Kleugel et al. (1977), objective measures of income and occupational prestige drive a latent factor, Subjective SES, which in turn is related to subjective measures of occupational prestige, income and overall social status. The latter three variables are considered measured with error, as is the latent factor Subjective SES. Graphically:



This model may be fit with the `sem` command as:

```
sem (SubjSES -> s_income s_occpres s_socstat) ///  
(SubjSES <- income occpres)
```

This model has both a structural component (relating the objective measures to Subjective SES) and a measurement component linking that factor to the three subjective measures.

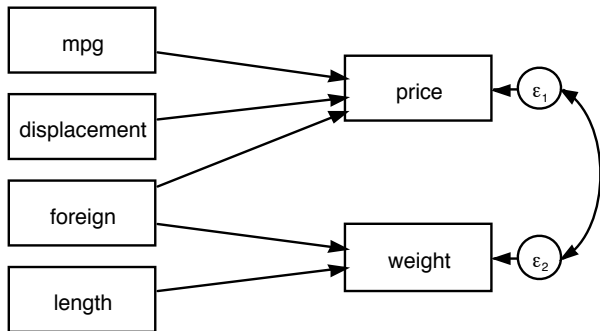
Seemingly unrelated regression

Seemingly unrelated regression is subsumed in the SEM framework as a pure structural model, with no measurement component nor latent variables. It could be fit with Stata's `sureg` command, or with `sem`:

```
sem (price <- foreign mpg displacement) ///  
    (weight <- foreign length), cov(e.price*e.weight)
```

which may be related to the graphical representation.

Here we have four exogenous variables and two continuous outcome variables. Their Gaussian error terms are assumed to be correlated. In the estimation, we may evaluate the strength of that correlation.



Latent growth model

The SEM framework is used to estimate latent growth models, where we are trying to evaluate the trajectory that a latent variable takes on over time. Using data from Bollen and Curran (2006), we have information on crime rates over several two-month periods for 369 communities. We may implement this model using `sem` as:

```
sem (lncrime0 <- Intercept@1 Slope@0 _cons@0) ///  
(lncrime1 <- Intercept@1 Slope@1 _cons@0) ///  
(lncrime2 <- Intercept@1 Slope@2 _cons@0) ///  
(lncrime3 <- Intercept@1 Slope@3 _cons@0), ///  
latent(Intercept Slope) ///  
var(e.lncrime0@var e.lncrime1@var ///  
e.lncrime2@var e.lncrime3@var) ///  
means(Intercept Slope)
```

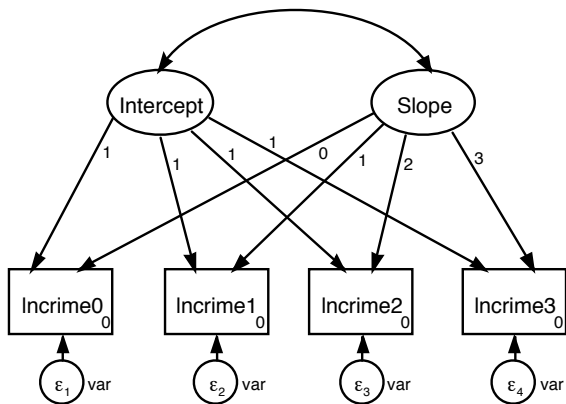
In this measurement model framework, we have four endogenous variables, the observed (log) crime rates, and two latent exogenous variables: the Intercept and Slope of the growth curves, such that

$$\text{Incrime}_\tau = \text{Intercept} + \tau \text{Slope}$$

Constraining the Slope coefficients to be 0, 1, 2, 3 imposes a linear growth curve. This could also be considered as a mixed model:

```
generate id = _n
reshape long lncrime, i(id) j(year)
mixed lncrime year || id:year, cov(unstructured) mle var
```

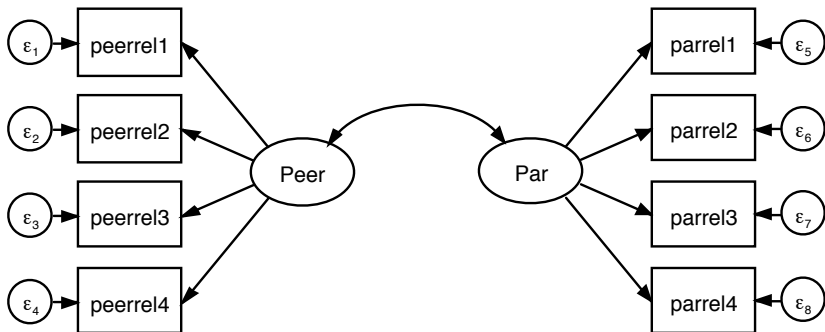
where Intercept and Slope are what would be called the fixed-effects coefficients in mixed.



Two-factor measurement model by group

We often want to test whether the same parameters apply to different groups in the data. The `sem` command has options that allow for differences in the path coefficients and covariances across groups of the data, such as males and females, or blacks, whites and Hispanics.

We consider the same sort of two-factor measurement model as in a prior example, using data for two groups of survey respondents: grade 4 and grade 5 students. Graphically:



The two latent factors are relationship with Peers and relationship with Parents. There are four measures available from the data for each latent factor. We may implement this model using `sem` as:

```
sem (Peer -> peerrel1 peerrel2 peerrel3 peerrel4) ///  
(Par -> parrel1 parrel2 parrel3 parrel4), group(grade)
```

The `group(grade)` option tells Stata that some of the model's parameters are to be constrained across the two groups, while others (e.g., the variances of each observed measurement) are estimated separately for each group.