Title

intro 4 — Tour of models

Description

Below is a sampling of structural equation models that can be fit by `sem`.

Remarks

If you have not read [SEM] **intro 2**, please do so. You need to speak the language. We also recommend reading [SEM] **intro 3**, but that is not required.

Now that you speak the language, we can start all over again and take a look at some of the classic models that `sem` can fit.

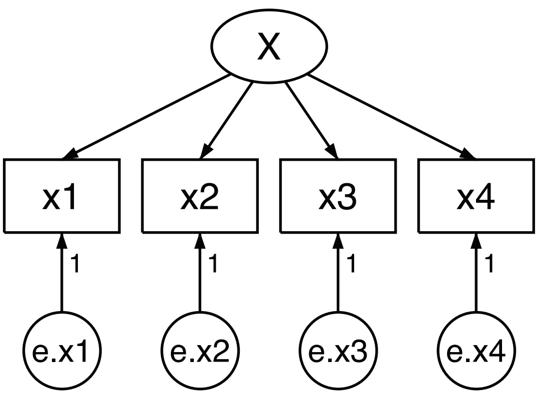
Remarks are presented under the following headings:

Single-factor measurement models
Multiple-factor measurement models
CFA models
Structural models 1: Linear regression
Structural models 2: Dependencies between endogenous variables
Structural models 3: Unobserved inputs, outputs, or both
Structural models 4: MIMIC
Structural models 5: Seemingly unrelated regression (SUR)
Structural models 6: Multivariate regression
Correlations
Higher-order CFA models
Correlated uniqueness model
Latent growth models
Models with reliability

Single-factor measurement models

See [SEM] **example 1**.

A single-factor measurement model is



The model can be written in Stata command language as

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

or as

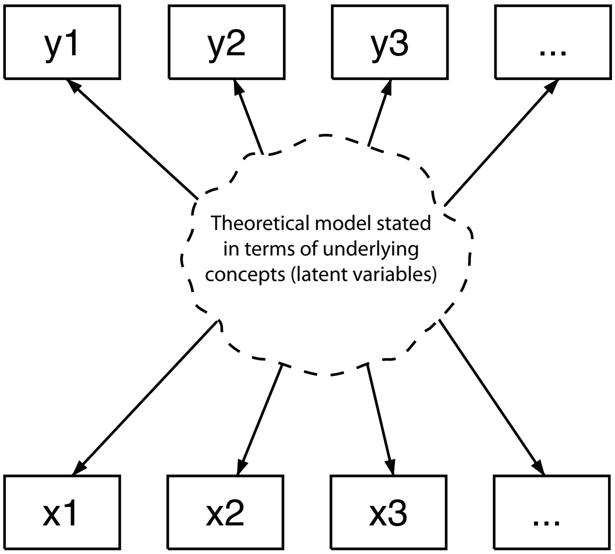
```
(x1 x2 x3 x4<-X)
```

or as

```
(X->x1 x2 x3 x4)
```

or in other ways. All the equivalent ways really are equivalent; no subtle differences will subsequently arise according to your choice.

The measurement model plays an important role in many other SEMs dealing with the observed inputs and the observed outputs:



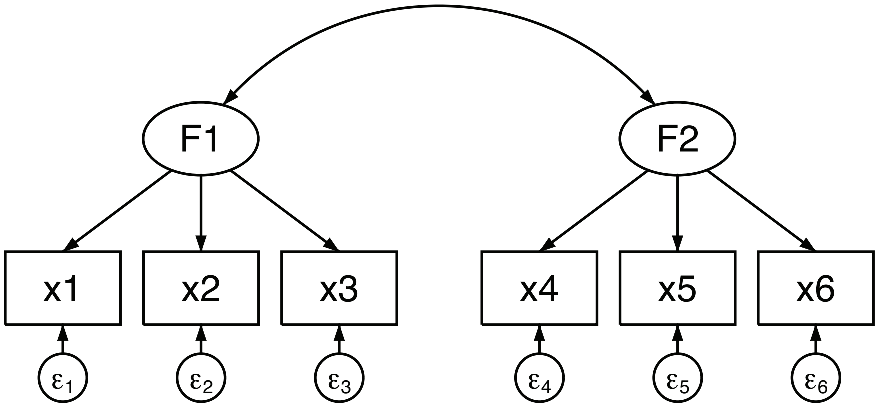
Because the measurement model is so often joined with other models, it is common to refer to the coefficients on the paths from latent variables to observable endogenous variables as the measurement coefficients and their intercepts as the measurement intercepts. The intercepts are usually not shown in path diagrams. The other coefficients and intercepts are those not related to the measurement issue.

The measurement coefficients are often referred to as loadings.

Multiple-factor measurement models

See [SEM] **example 3**.

A two-factor measurement model is two one-factor measurement models with possible correlation between the factors:



To obtain a correlation between F1 and F2, we drew a curved path.

The model can be written in Stata command language as

```
(F1->x1) (F1->x2) (F1->x3) (F2->x4) (F2->x5) (F2->x6)
```

In the command language, we do not have to include the `cov(F1*F2)` option because, by default, `sem` assumes that exogenous latent variables are correlated with each other.

This model can also be written in any of the following ways:

```
(F1->x1 x2 x3) (F2->x4 x5 x6)
```

or

```
(x1 x2 x3<-F1) (x4 x5 x6<-F2)
```

or

```
(x1<-F1) (x2<-F1) (x3<-F1) (x4<-F2) (x5<-F2) (x6<-F2)
```

CFA models

See [SEM] **example 5**.

The measurement models just shown are also known as confirmatory factor analysis (CFA) models because they can be analyzed using CFA.

In the single-factor model, after estimation, you might want to test that all the indicators have significant loadings by using `test`; see [SEM] **test**. You might also want to test whether the correlations between the errors should have been included in the model by using `estat mindices`; see [SEM] **estat mindices**.

In the multiple-factor measurement model, you might want to test that any of the omitted paths should in fact be included in the model. The omitted paths in the two-factor measurement model above were F1 → x4, F1 → x5, F1 → x6, and F2 → x1, F2 → x2, F2 → x3. `estat mindices` will perform these tests.

We show other types of CFA models below.

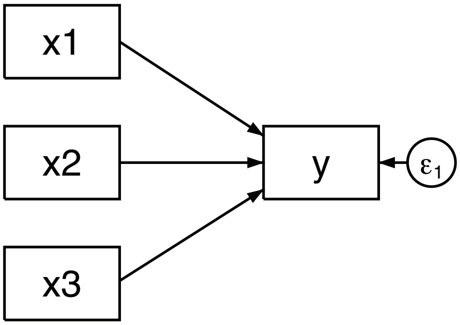
Structural models 1: Linear regression

See [SEM] **example 6**.

Different authors define the meaning of structural models in different ways. Bollen (1989, 4) defines a structural model as the parameters being not just of a descriptive nature of association but instead of a casual nature. By that definition, the measurement models above could be structural models, and so could the linear regression below.

Others define structural models as models having paths reflecting causal dependencies between endogenous variables and thus would exclude the measurement model and linear regression. We will show you a "true" structural model in the next example.

An example of a linear regression would be



The model above can be written in Stata command language as

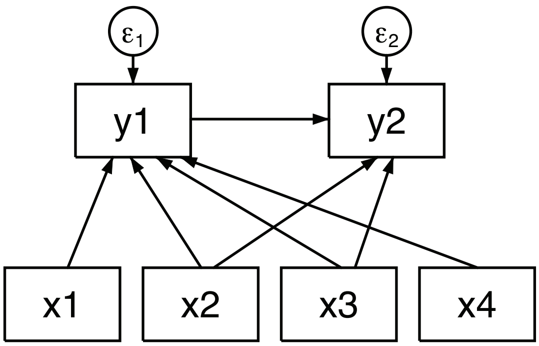
```
(y <- x1 x2 x3)
```

When you estimate a linear regression using `sem`, you obtain the same point estimates as you would with `regress` and the same standard errors up to a degree-of-freedom adjustment applied by `regress`.

Structural models 2: Dependencies between endogenous variables

See [SEM] **example 7**.

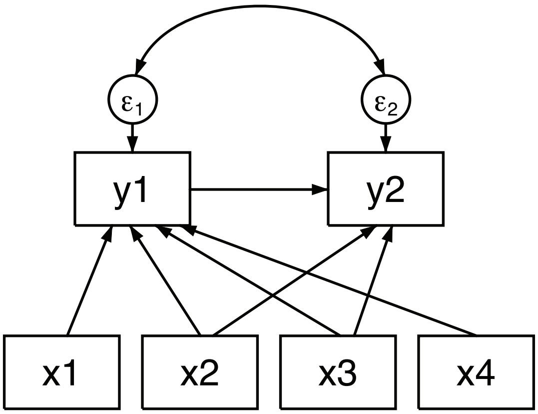
An example of a structural model having paths between endogenous variables would be



The model above can be written in Stata command language as

```
(y1 <- x1 x2 x3 x4) (y2 <- y1 x2 x3)
```

In this example, all inputs and outputs are observed and the errors are assumed to be uncorrelated. In these kinds of models, it is common to allow correlation between errors:



The model above can be written in Stata command language as

```
(y1 <- x1 x2 x3 x4) (y2 <- y1 x2 x3), cov(e.y1*e.y2)
```

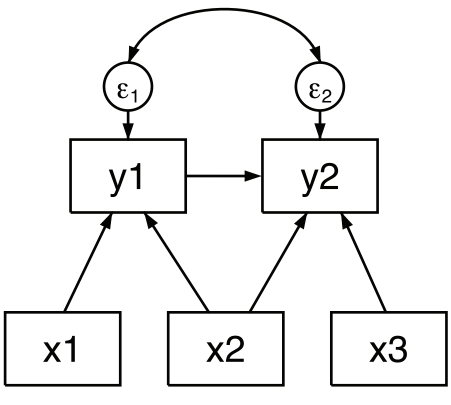
This structural model is said to be overidentified. If we omitted `y1` ← `x4`, the model would be just identified. If we also omitted `y1` ← `x1`, the model would be unidentified.

When you fit the above model using `sem`, you obtain slightly different results from those you would obtain with `ivregress liml`. This is because `sem` with default `method(ml)` produces full information maximum likelihood rather than limited-information maximum likelihood results.

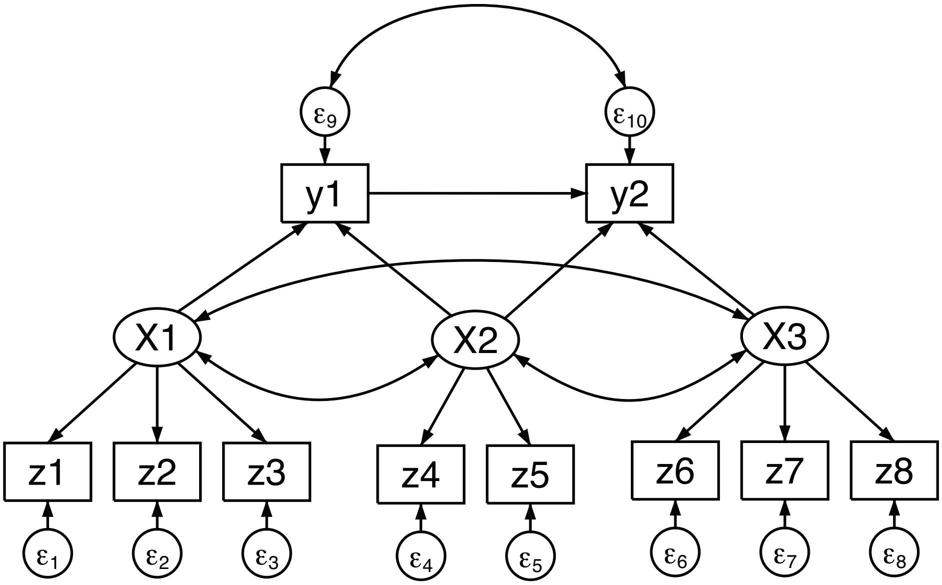
Structural models 3: Unobserved inputs, outputs, or both

See [SEM] **example 9**.

Perhaps in a structural model such as



the inputs x1, x2, and x3 are concepts and thus are not observed. Assume that we have measurements for them. We can join this structural model example with a three-factor measurement model:



Note the curved arrows denoting correlation between the pairs of X1, X2, and X3. In the previous path diagram, we had no such arrows between the variables, yet we were still assuming that they were there. In sem's path diagrams, correlation between exogenous observed variables is assumed and need not be explicitly shown. When we changed observed variables x1, x2, and x3 to be the latent variables X1, X2, and X3, we needed to show explicitly the correlations we were allowing. Correlation between latent variables is not assumed unless shown.

This model can be written in Stata command syntax as follows:

```
(y1<-X1 X2) (y2<-y1 X2 X3)    ///
(X1->z1 z2 z3)                ///
(X2->z4 z5)                   ///
(X3->z6 z7 z8),               ///
          cov(e.y1*e.y2)
```

We did not include the cov(X1*X2 X1*X3 X2*X3) option, although we could have. In the command language, exogenous latent variables are assumed to be correlated with each other. If we did not want X2 and X3 to be correlated, we would need to include the cov(X2*X3@0) option.

We changed x1, x2, and x3 to be X1, X2, and X3. In command syntax, variables beginning with a capital letter are assumed to be latent. Alternatively, we could have left the names in lowercase and specified the identities of the latent variables:

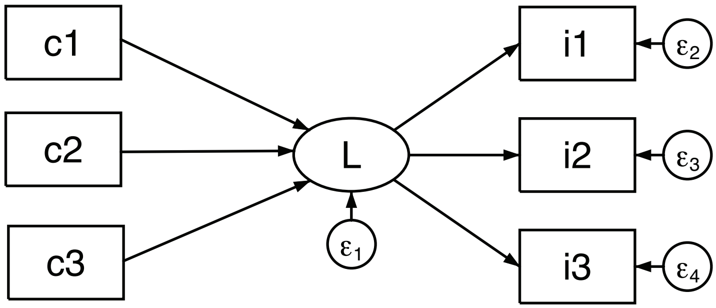
```
(y1<-x1 x2) (y2<-y1 x2 x3)    ///
(x1->z1 z2 z3)                ///
(x2->z4 z5)                   ///
(x3->z6 z7 z8),               ///
          cov(e.y1*e.y2)      ///
          latent(x1 x2 x3)
```

Just as we have joined an observed structural model to a measurement model to handle unobserved inputs, we could join the above model to a measurement model to handle unobserved y1 and y2.

Structural models 4: MIMIC

See [SEM] **example 10**.

MIMIC stands for multiple indicators and multiple causes. An example of a MIMIC model is



In this model, the observed causes c1, c2, and c3 determine latent variable L, and L in turn determines observed indicators i1, i2, and i3.

This model can be written in Stata command syntax as

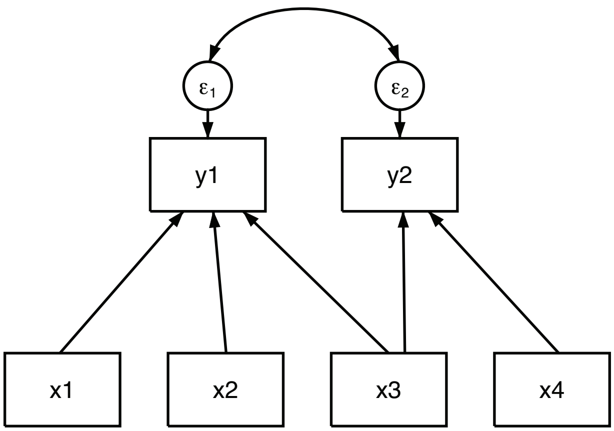
```
(i1 i2 i3 <- L) (L <- c1 c2 c3)
```

Structural models 5: Seemingly unrelated regression (SUR)

See [SEM] **example 12**.

Seemingly unrelated regression is like having two or more separate linear regressions but allowing the errors to be correlated.

An example of a seemingly unrelated regression (SUR) model is



The model above can be written in Stata command syntax as

```
(y1 <- x1 x2 x3) (y2 <- x3 x4), cov(e.y1*e.y2)
```

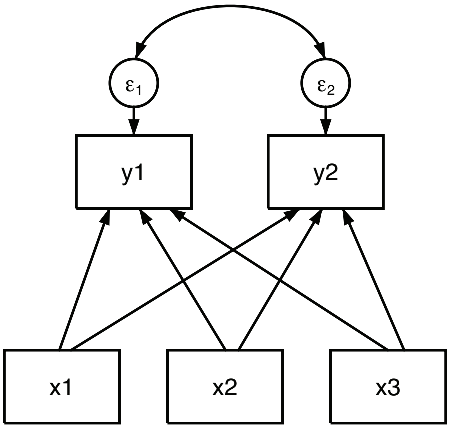
In this example, the two regressions shared a common exogenous variable, x3. That is not necessary. Or, they could share more variables. If they shared all variables, results would be the same as estimating multivariate regression, shown in the next example.

When you estimate an SUR using sem, you obtain the same point estimates as you would with sureg if you specify sureg's isure option, which causes sureg to iterate until it obtains the maximum likelihood result. Standard errors will be different. If the model has exogenous variables only on the right-hand side, standard errors will be asymptotically identical and, although the standard errors are different in finite samples, there is no reason to prefer one set over the other. If the model being fit is recursive, standard errors produced by sem are better than those from sureg, both asymptotically and in finite samples.

Structural models 6: Multivariate regression

See [SEM] **example 12**, even though the example is of SUR. Multivariate regression is a special case of SUR.

A multivariate regression is just an SUR where the different dependent variables share the same exogenous variables:



The model above can be written in Stata command syntax as

```
(y1 y2 <- x1 x2 x3), cov(e.y1*e.y2)
```

When you estimate a multivariate regression using sem, you obtain the same point estimates as you would with mvreg and the same standard errors up to a multiplicative $\sqrt{(N-p-1)/N}$ degree-of-freedom adjustment applied by mvreg.

Correlations

See [SEM] **example 16**.

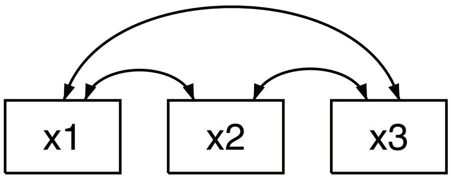
We are all familiar with correlation matrices of observed variables, such as

	x1	x2	x3
x1	1.0000		
x2	0.7700	1.0000	
x3	−0.0177	−0.2229	1.0000

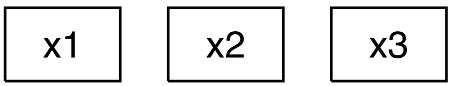
or covariances matrices, such as

	x1	x2	x3
x1	662.172		
x2	62.5157	9.95558	
x3	−0.769312	−1.19118	2.86775

These results can be obtained from `sem`. The path diagram for the model is



We could just as well leave off the curved paths because `sem` assumes them among observed exogenous variables:



Either way, this model can be written in Stata command syntax as

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

That is, we simply omit specifying the target of the path, the endogenous variable.

If we fit the model, we will obtain the covariance matrix by default. `correlate` with the `covariance` option produces covariances that are divided by $N-1$ rather than N. To match this covariance exactly, you need to specify the `nm1` option, which we can do in the command language by typing

```
(<- x1 x2 x3), nm1
```

If we want correlations rather than covariances, we ask for them by specifying the `standardized` option:

```
(<- x1 x2 x3), nm1 standardized
```

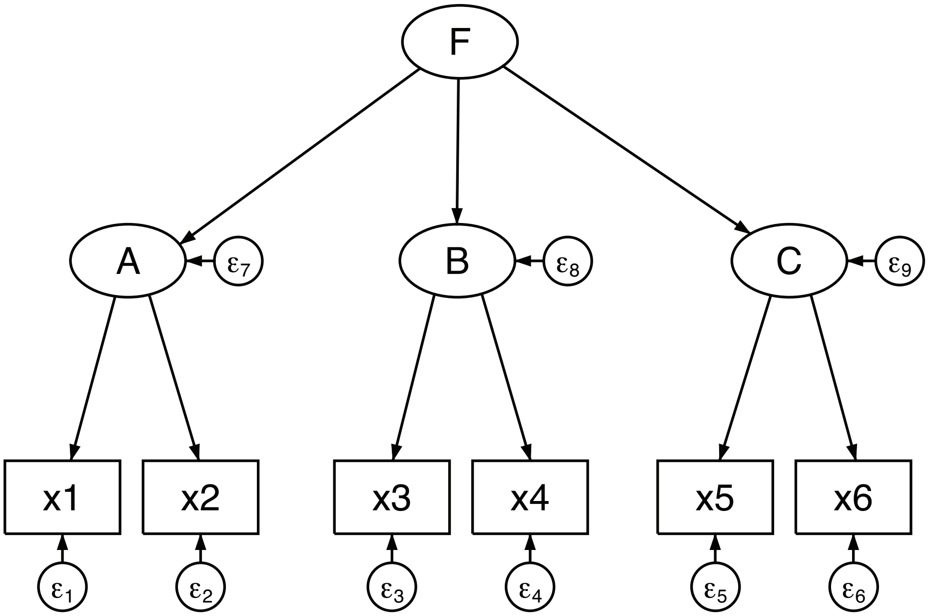
An advantage of obtaining correlation matrices from sem rather than from correlate is that you can perform statistical tests on the results, such as that the correlation of x1 and x3 is equal to the correlation of x2 and x3.

If you are willing to assume joint normality of the variables, you can obtain more efficient estimates of the correlations in the presence of missing-at-random data by specifying the method(mlmv) option.

Higher-order CFA models

See [SEM] **example 15**.

Sometimes observed values measure traits or other aspects of latent variables, so we insert a new layer of latent variables to reflect those traits or aspects. We have measurements—say, x1, ..., x6—all reflecting underlying factor F, but x1 and x2 measure one trait of F, x3 and x4 measure another trait, and x5 and x6 measure yet another trait. This model can be drawn as



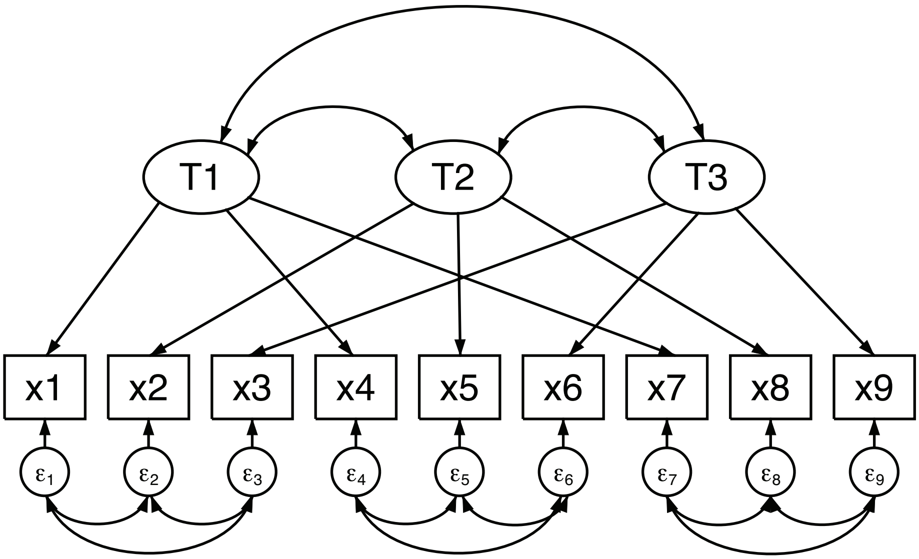
The above model can be written in command syntax as

```
(A->x1 x2) (B->x3 x4) (C->x5 x6) (A B C <- F)
```

Correlated uniqueness model

See [SEM] **example 17**.

Sometimes observed values are correlated just because of how the data are collected. Imagine we have factor T1 representing a trait and measurements x1 and x4. Perhaps T1 is aggression, and x1 is self reported, and x4 is reported by the spouse. Imagine we also have another factor, T2, and measurements x2 and x5; x2 is self reported, and x5 is reported by the spouse. It would not be unlikely that x1 and x2 are correlated and that x4 and x5 are correlated. That is exactly what the correlated uniqueness model assumes:



Data that exhibit this kind of pattern are known as multitrait–multimethod (MTMM) data. Researchers historically looked at the correlations, but SEM allows us to fit a model that incorporates the correlations.

The above model can be written in Stata command syntax as

```
(T1 -> x1 x4 x7)                                          ///
(T2 -> x2 x5 x8)                                          ///
(T3 -> x3 x6 x9),                                         ///
             cov(e.x1*e.x2  e.x1*e.x3  e.x2*e.x3)         ///
             cov(e.x4*e.x5  e.x4*e.x6  e.x5*e.x6)         ///
             cov(e.x7*e*x8  e.x7*e.x9  e.x8*e.x9)
```

An alternative way to type the above is to use the `covstructure()` option, which we can abbreviate as `covstruct()`:

```
(T1 -> x1 x4 x7)                                           ///
(T2 -> x2 x5 x8)                                           ///
(T3 -> x3 x6 x9),                                          ///
             covstruct(e.x1 e.x2 e.x3, unstructured)       ///
             covstruct(e.x4 e.x5 e.x6, unstructured)       ///
             covstruct(e.x7 e.x8 e.x9, unstructured)
```

Unstructured means that the listed variables have covariances. Specifying blocks of errors as unstructured would save typing if there were more variables in each block.

Latent growth models

See [SEM] **example 18**.

A latent growth model is a variation on the measurement model. In our measurement model examples, we have assumed four observed measurements of underlying factor `X`: `x1`, `x2`, `x3`, and `x4`. In the command language, which saves paper, we can write this as

```
(X->x1) (X->x2) (X->x3) (X->x4)
```

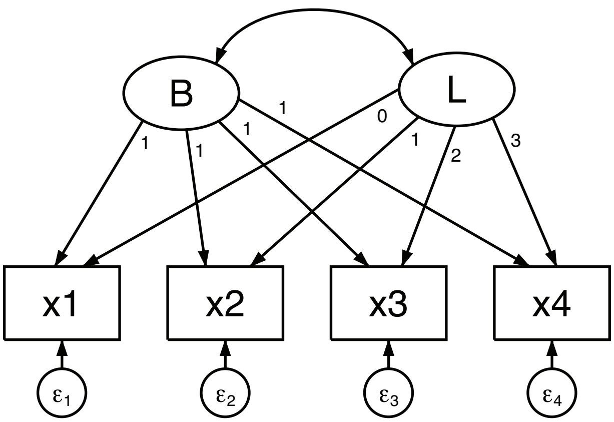
Let's assume that the observed values are collected over time. x1 is observed at time 0, x2 at time 1, and so on. It thus may be more reasonable to assume that the observed values represent a base value and a growth modeled with a linear trend. Thus we might write the model as

```
(B@1 L@0 -> x1)              ///
(B@1 L@1 -> x2)              ///
(B@1 L@2 -> x3)              ///
(B@1 L@3 -> x4),             ///
          noconstant
```

which is to say, the equations are

$$x_1 = B + 0L + e.x_1$$
$$x_2 = B + 1L + e.x_1$$
$$x_3 = B + 2L + e.x_1$$
$$x_4 = B + 3L + e.x_1$$

The path diagram for the model is



In evaluating this model, it is useful to review the means of the latent exogenous variables. In most models, latent exogenous variables have mean 0 and the means are thus uninteresting. sem usually constrains latent exogenous variables to have mean 0 and does not report that fact.

In this case, however, we ourselves have placed constraints, and thus the means are identified and in fact are an important point of the exercise. We must tell sem not to constrain the means of the two latent exogenous variables B and L, which we do with the means() option:

```
(B@1 L@0 -> x1)              ///
(B@1 L@1 -> x2)              ///
(B@1 L@2 -> x3)              ///
(B@1 L@3 -> x4),             ///
          noconstant means(B L)
```

We must similarly specify the means() option when using the GUI.

Models with reliability

See [SEM] **example 24**.

A typical solution for dealing with variables measured with error is to find multiple measurements and to use those measurements to develop a latent variable. See, for example, *Single-factor measurement models* and *Multiple-factor measurement models* above.

When the reliability of the variables is known—reliability is measured as the fraction of variances that is not due to measurement error—another approach is available. This approach can be used in place of or in addition to the use of multiple measurements.

See [SEM] **sem option reliability( )**.

Also see

[SEM] **intro 3** — Substantive concepts

[SEM] **intro 6** — Postestimation tests and predictions

[SEM] **example 1** — Single-factor measurement model