Description Remarks and examples References Also see

Description

Below is a sampling of SEMs that can be fit by `sem` or `gsem`.

Remarks and examples

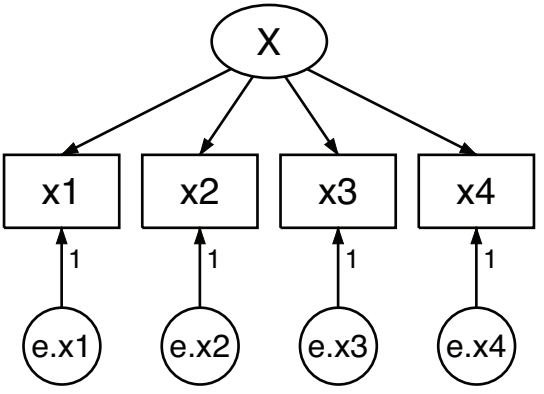
If you have not read [SEM] **Intro 2**, please do so. You need to speak the language. We also recommend reading [SEM] **Intro 4**, but that is not required.

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

Remarks are presented under the following headings:

Single-factor measurement models

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

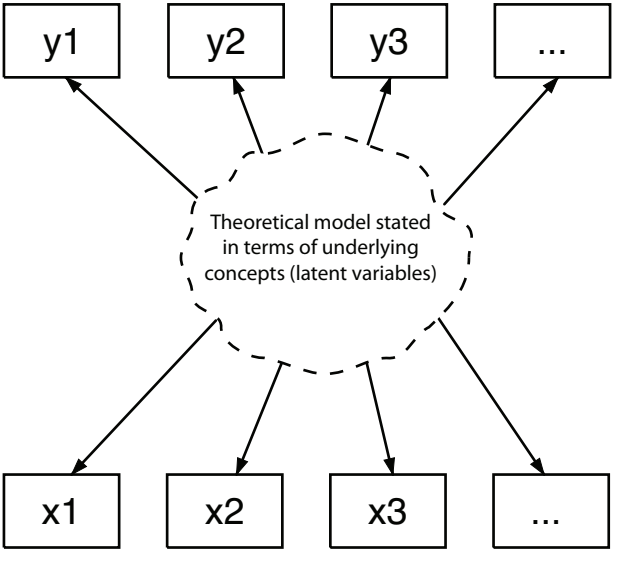
```
(x1 x2 x3 x4<-X)
```

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 to refer to 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.

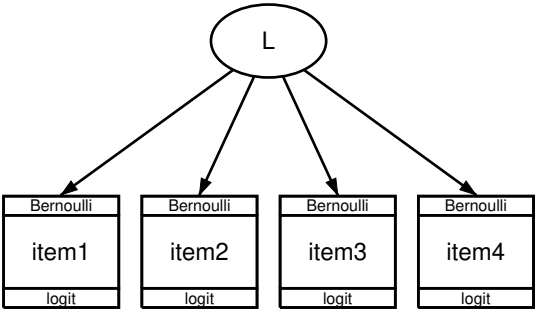
This model can be fit by `sem` or `gsem`. Use `sem` for standard linear models (standard means single level); use `gsem` when you are fitting a multilevel model or when the response variables are generalized linear such as probit, logit, multinomial logit, Poisson, and so on.

See the following examples:

1. [SEM] **Example 1**. Single-factor measurement model.
2. [SEM] **Example 27g**. Single-factor measurement model (generalized response).
3. [SEM] **Example 30g**. Two-level measurement model (multilevel, generalized response).
4. [SEM] **Example 35g**. Ordered probit and ordered logit.

Item response theory (IRT) models

Item response theory (IRT) models look like the following:



The items are the observed variables, and each has a 0/1 outcome measuring a latent variable. Often, the latent variable represents ability. These days, it is traditional to fit IRT models using logistic regression, but in the past, probit was used and they were called normal ogive models.

In one-parameter logistic models, also known as 1-PL models and Rasch models, constraints are placed on the paths and perhaps the variance of the latent variable. Either path coefficients are constrained to 1 or path coefficients are constrained to be equal and the variance of the latent variable is constrained to be 1. Either way, this results in the negative of the intercepts of the fitted model being a measure of difficulty.

1-PL and Rasch models can be fit treating the latent variable—ability—as either fixed or random. Abilities are treated as random with `gsem`.

In two-parameter logistic models (2-PL), no constraints are imposed beyond the one required to identify the latent variable, which is usually done by constraining the variance to 1. This results in path coefficients measuring discriminating ability of the items, and difficulty is measured by the negative of the intercepts divided by the corresponding (slope) coefficient.

IRT has been extended beyond 1-PL and 2-PL models, including extension to other types of generalized responses.

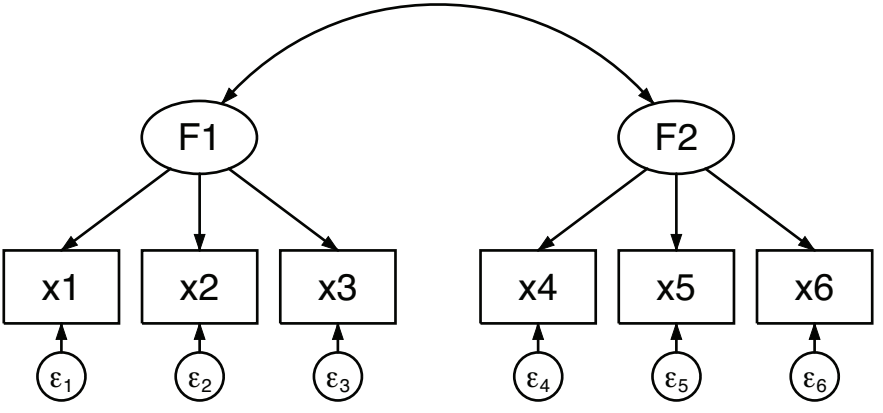
IRT models, including the extensions, can be fit by `gsem`.

See the following examples:

1. [SEM] **Example 28g**. One-parameter logistic IRT (Rasch) model.
2. [SEM] **Example 29g**. Two-parameter logistic IRT model.

Multiple-factor measurement models

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)
```

Two-factor measurement models can be fit by `sem` and `gsem`.

See the following examples:

1. [SEM] **Example 3**. Two-factor measurement model.
2. [SEM] **Example 31g**. Two-factor measurement model (generalized response).

Confirmatory factor analysis (CFA) models

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.

These tests are available after `sem` only. CFA is just a measurement model and can be fit by both `sem` and `gsem`.

We show other types of CFA models below.

See the following example:

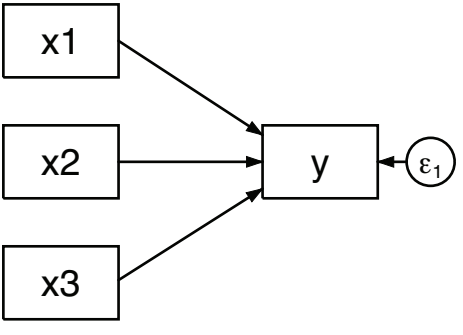
1. [SEM] **Example 5**. Modification indices.

Structural models 1: Linear regression

Different authors define the meaning of structural models in different ways. Bollen (1989, 4) defines a structural model as the parameters being not of a descriptive nature of association but instead of a causal 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, which would thus 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 is



This model can be written in Stata command language as

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

When you estimate a linear regression by 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`.

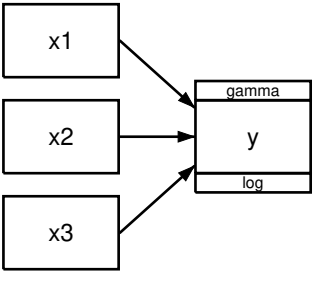
Linear regression models can be fit by `sem` and `gsem`. `gsem` also has options for censoring.

See the following examples:

1. [SEM] **Example 6**. Linear regression.
2. [SEM] **Example 38g**. Random-intercept and random-slope models (multilevel).
3. [SEM] **Example 40g**. Crossed models (multilevel).
4. [SEM] **Example 43g**. Tobit regression.
5. [SEM] **Example 44g**. Interval regression.

Structural models 2: Gamma regression

Gamma regression, also known as log-gamma regression, is used when a continuous outcome is nonnegative, when it ranges from zero to infinity, and often with positively skewed data. It is appropriate when the error variance can be assumed to increase with the mean. Gamma regression gives similar results to linear regression with a logged dependent variable.



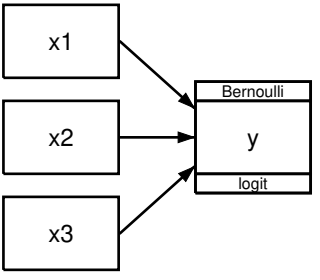
Gamma regression is fit by `gsem`; specify shorthand `gamma` or specify `family(gamma) link(log)`.

You can fit exponential regressions using gamma regression if you constrain the log of the scale parameter to be 0; see [SEM] **gsem family-and-link options**.

Structural models 3: Binary-outcome models

Binary-outcome models have 0/1 response variables. These models include logistic regression (also known as logit), probit, and complementary log–log (also known as cloglog) models.

A simple logistic regression model is



which in command syntax can be written as

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

For the other binary-outcome models, all that changes in the diagram is the names of the family and link; in the command language, the option name changes; see [SEM] **gsem family-and-link options**.

Usually, the observations in binary-outcome data record whether the event occurred, but the data can instead record the number of events and the number of trials by changing the family from Bernoulli to binomial; see [SEM] **gsem family-and-link options**.

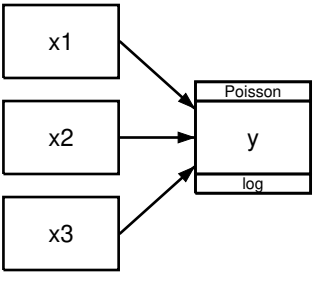
Binary-outcome models can be fit by `gsem`.

See the following examples:

1. [SEM] **Example 33g**. Logistic regression.
2. [SEM] **Example 27g**. Single-factor measurement model (generalized response).
3. [SEM] **Example 34g**. Combined models (generalized responses).

Structural models 4: Count models

Count models have response variables that are counts of things or events. These models include Poisson and negative binomial models. A simple example of a Poisson model is



which in command syntax can be written as

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

For negative binomial models, in path diagrams, the family changes to nbreg and, in the command syntax, the option changes to `nbreg`.

In Poisson models, both the mean and the variance are determined by a single parameter: the rate at which the event occurs. One use of negative binomial models is to handle overdispersion, which is when the variance is greater than the variance that would be predicted by a Poisson model.

The way we have diagrammed the model and written the command, we are assuming that each observation was at risk for the same length of time. If observations varied in this, the command would be

```
(y<-x1 x2 x3, poisson exposure(etime))
```

where variable `etime` records each observation's exposure time. The diagram would not change, but we would nonetheless enter the exposure time into the Builder. See [SEM] **gsem family-and-link options**.

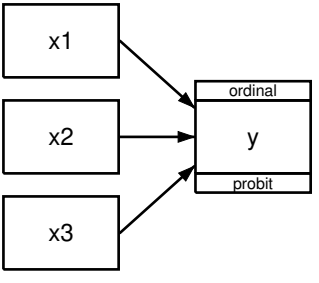
Count models are fit by `gsem`.

See the following examples:

1. [SEM] **Example 34g**. Combined models (generalized responses).
2. [SEM] **Example 39g**. Three-level model (multilevel, generalized response).

Structural models 5: Ordinal models

Ordinal models have two or more possible outcomes that are ordered, such as responses of the form "a little", "average", and "a lot"; or "strongly disagree", "disagree", ..., "strongly agree". The outcomes are usually numbered 1, 2, ..., k. These models include ordered probit, ordered logit, and ordered complementary log–log (also known as ocloglog). A simple example of an ordered probit model is



which in command syntax can be written as

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

All that changes for the other models is the link name that appears in the path diagram or the option that appears in the command language.

Ordinal models are fit by `gsem`.

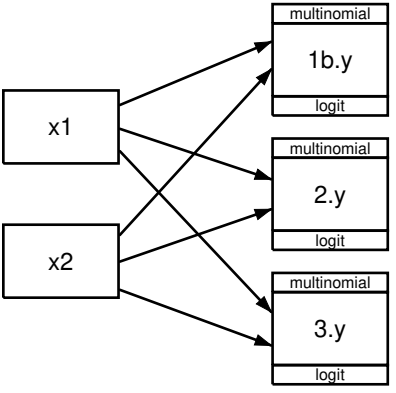
See the following examples:

1. [SEM] **Example 35g**. Ordered probit and ordered logit.
2. [SEM] **Example 31g**. Two-factor measurement model (generalized response).
3. [SEM] **Example 32g**. Full structural equation model (generalized response).
4. [SEM] **Example 36g**. MIMIC model (generalized response).

Structural models 6: Multinomial logistic regression

Multinomial logistic regression, also known as multinomial logit, is similar to ordinal models in that it, too, deals with multiple responses; however, in multinomial logistic regression, the responses cannot be ordered. Examples of multinomial logit response include method of transportation (car, public transportation, etc.) or ice-cream flavor (vanilla, chocolate, etc.).

Just as with ordinal models, the outcome is usually recorded as a single variable containing 1, 2, …, k, but in path diagrams, factor-variable notation is used to identify the outcomes:



In command syntax, this can be written as

```
(i.y<-x1 x2, mlogit)
```

In the example above, we have paths from all predictor variables to all outcomes. That is common but not required.

The multinomial logistic regression model is fit by `gsem`.

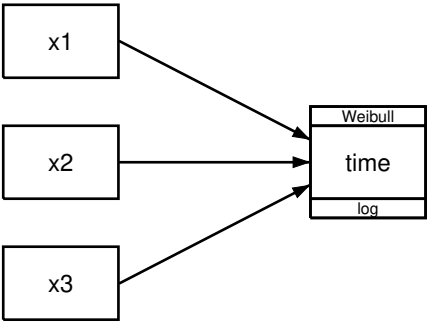
See the following examples:

1. [SEM] **Example 37g**. Multinomial logistic regression.
2. [SEM] **Example 41g**. Two-level multinomial logistic regression (multilevel).

Structural models 7: Survival models

Survival models are fit to response variables that measure survival times. The data may also contain observations for which a failure is not observed. For those observations, our response variable represents the time until the observation is censored.

Parametric survival models may be fit using a variety of distributions. A simple example of a Weibull survival model is



which in command syntax can be written as

```
(time<-x1 x2 x3, family(weibull))
```

If some of the observations were censored, the command would be

```
(time<-x1 x2 x3, family(weibull, failure(fail)))
```

where fail is an indicator variable coded as 1 for observations that failed and 0 for observations that were censored. The diagram would not change, but we would nonetheless specify fail as the failure indicator by using the Builder.

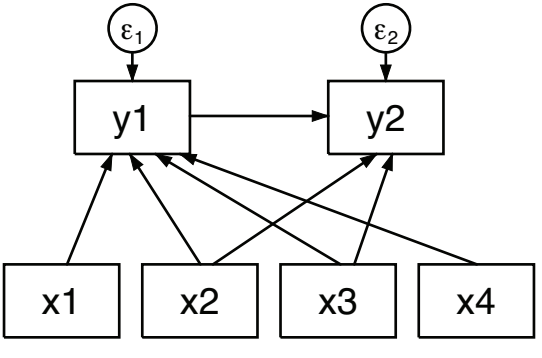
`gsem` fits parametric survival models using exponential, Weibull, gamma, loglogistic, and lognormal distributions. See [SEM] **gsem family-and-link options** for details of all options available when fitting survival models, including those for specifying the survival distribution and censoring.

See the following examples:

1. [SEM] **Example 47g**. Exponential survival model.
2. [SEM] **Example 48g**. Loglogistic survival model with censored and truncated data.
3. [SEM] **Example 49g**. Multiple-group Weibull survival model.

Structural models 8: Dependencies between response variables

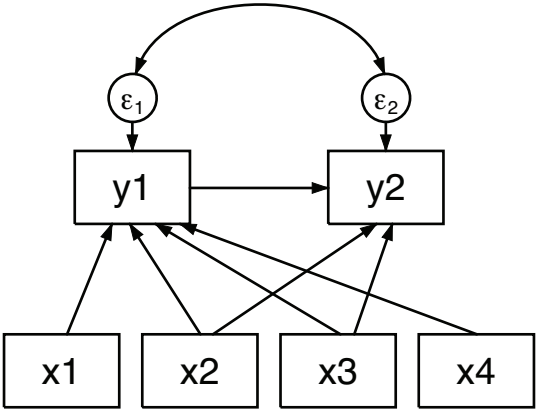
An example of a structural model having paths between response (endogenous) variables is



This model 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.

Analysis of models for observed variables that include dependencies between endogenous variables may also be referred to as path analysis. Acock (2013, chap. 2) discusses path analysis with `sem` in more detail.

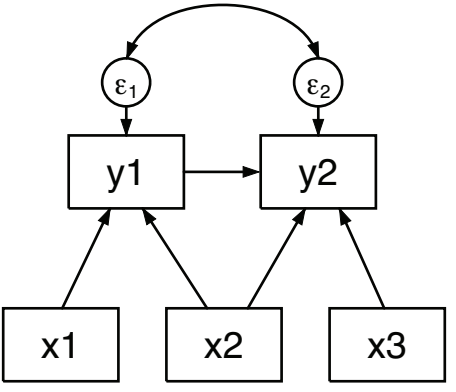
These models can be fit by `sem` or `gsem`. When using `gsem` to fit models with generalized response variables, non-Gaussian responses and Gaussian responses with the log link, or censoring, can only be included in recursive portions of models.

See the following examples:

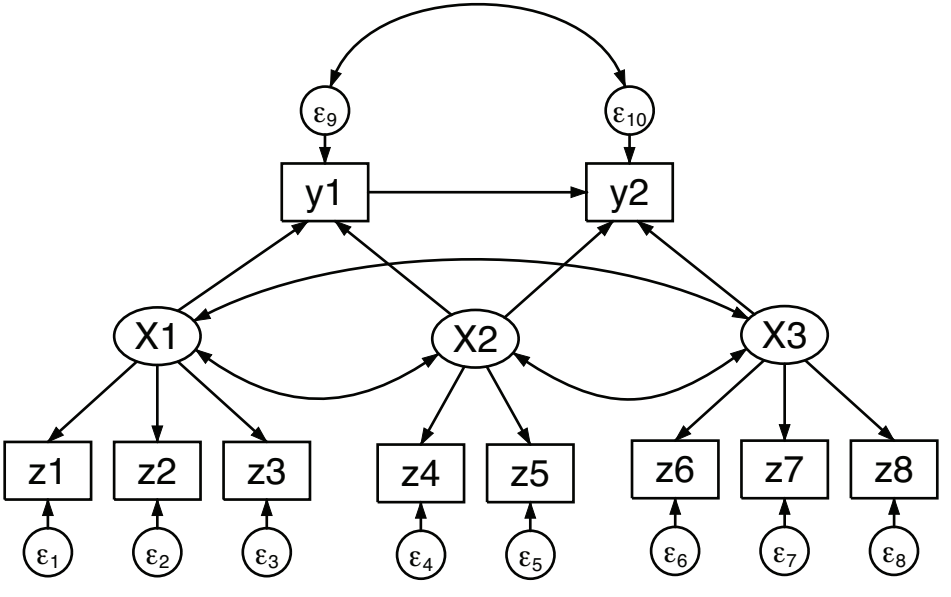
1. [SEM] **Example 7**. Nonrecursive structural model.
2. [SEM] **Example 34g**. Combined models (generalized responses).
3. [SEM] **Example 42g**. One- and two-level mediation models (multilevel).
4. [SEM] **Example 43g**. Tobit regression.
5. [SEM] **Example 44g**. Interval regression.
6. [SEM] **Example 45g**. Heckman selection model.
7. [SEM] **Example 46g**. Endogenous treatment-effects model.

Structural models 9: Unobserved inputs, outputs, or both

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 and must be 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`.

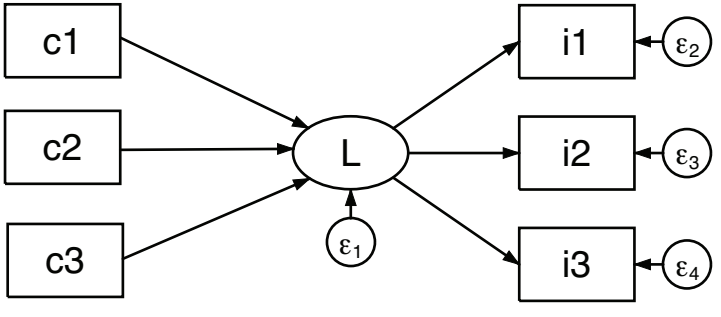
Models with unobserved inputs, outputs, or both can be fit by `sem` and `gsem`.

See the following examples:

1. [SEM] **Example 9**. Structural model with measurement component.
2. [SEM] **Example 32g**. Full structural equation model (generalized response).
4. [SEM] **Example 45g**. Heckman selection model.
5. [SEM] **Example 46g**. Endogenous treatment-effects model.

Structural models 10: MIMIC models

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 the observed indicators `i1`, `i2`, and `i3`.

This model can be written in Stata command syntax as

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

MIMIC models can be fit by `sem` and `gsem`.

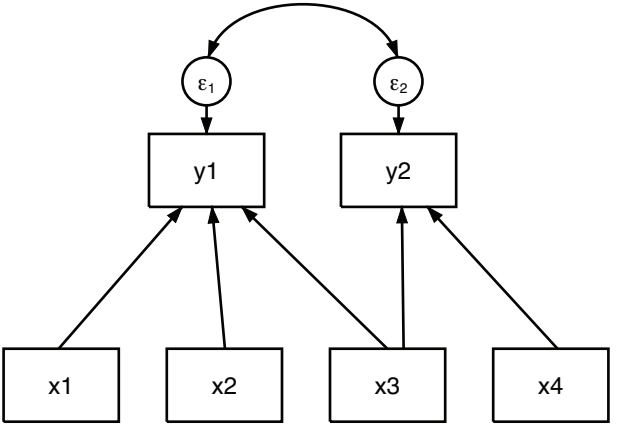
See the following examples:

1. [SEM] **Example 10**. MIMIC model.
2. [SEM] **Example 36g**. MIMIC model (generalized response).

Structural models 11: Seemingly unrelated regression (SUR)

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

An example of an 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 share a common exogenous variable, `x3`. They do not have to share a common variable, 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 with `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, then 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.

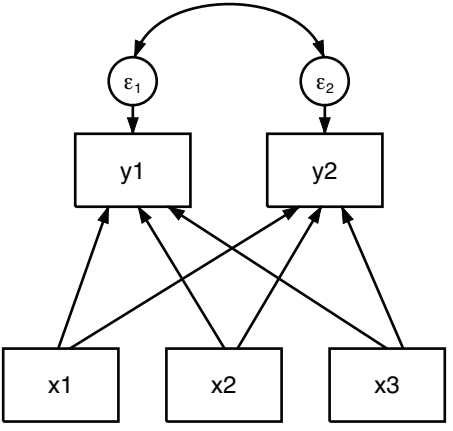
SUR models can be fit by `sem` and `gsem`. `gsem` will allow you to generalize the model to multilevel. In cases other than family Gaussian, link identity, you can sometimes introduce latent variables to create a similar effect.

See the following example:

1. [SEM] **Example 12**. Seemingly unrelated regression.

Structural models 12: Multivariate regression

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 with `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`.

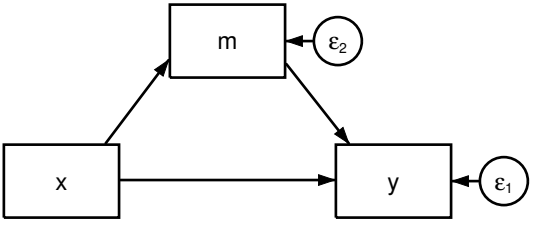
Multivariate regression can be fit by `sem` and `gsem`. `gsem` will allow you to generalize the model to multilevel. In cases other than family Gaussian, link identity, you can sometimes introduce latent variables to create a similar effect.

See the following example:

1. [SEM] **Example 12**. Seemingly unrelated regression.

Structural models 13: Mediation models

Mediation models concern effects that pass through (are mediated by) other variables. Consider response variable y affected by x mediated through m:



This can also be specified in command syntax as

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

In this simple model, x has a direct effect on y and an indirect (mediated through m) effect. The direct effect may be reasonable given the situation, or it may be included just so one can test whether the direct effect is present. If both the direct and indirect effects are significant, the effect of x is said to be partially mediated through m.

There are one-level mediation models and various two-level models, and lots of other variations, too.

`sem` and `gsem` can both fit one-level linear models, but you will be better off using `sem`. `gsem` can fit one-level generalized linear models and fit two-level (and higher) models, generalized linear or not.

See the following example:

1. [SEM] **Example 42g**. One- and two-level mediation models (multilevel).

Correlations

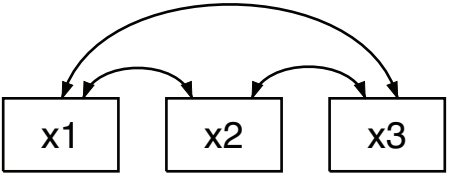
We are all familiar with correlation matrices of observed variables, such as

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

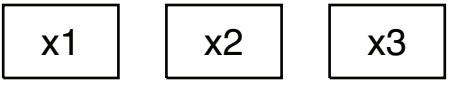
or covariances matrices, such as

	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 by 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.

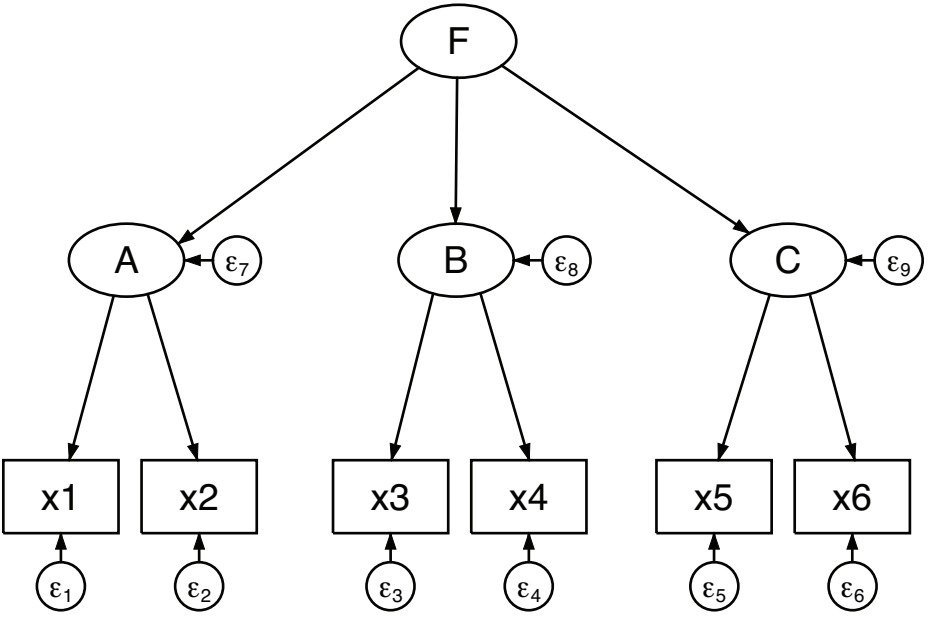
Correlations are fit using `sem`.

See the following example:

1. [SEM] **Example 16**. Correlation.

Higher-order CFA models

Observed values sometimes 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 model can be written in command syntax as

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

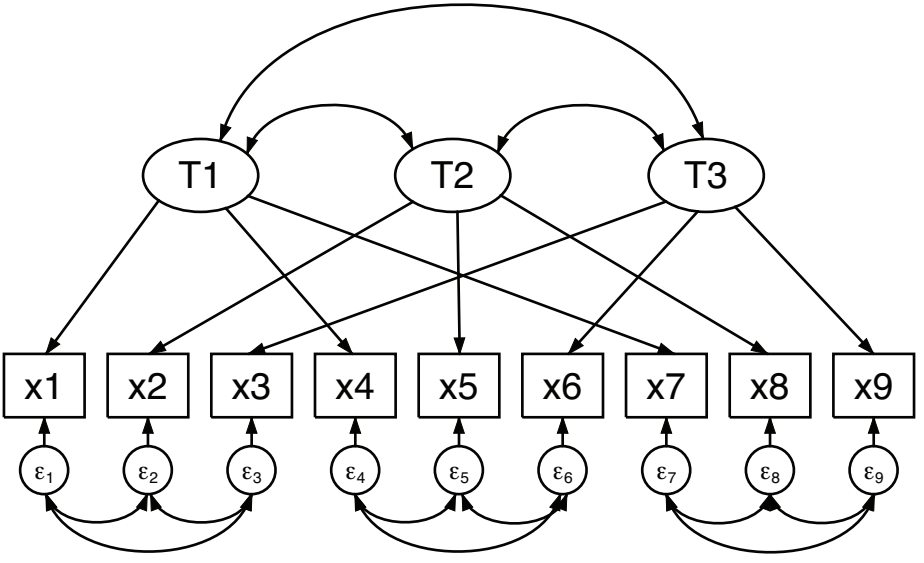
Higher-order CFA models can be fit using `sem` or `gsem`.

See the following example:

1. [SEM] **Example 15**. Higher-order CFA.

Correlated uniqueness model

Observed values sometimes are correlated just because of how the data are collected. Imagine we have factor `T1` representing a trait with measurements `x1` and `x4`. Perhaps `T1` is aggression, and then `x1` is self reported and `x4` is reported by the spouse. Imagine we also have factor `T2` with measurements `x2` and `x5`. Again, `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 structural equation modeling 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.

The correlated uniqueness model can be fit by `sem` or `gsem`, although we recommend use of `sem` in this case. Gaussian responses with the identity link are allowed to have correlated uniqueness (error) but only in the absence of censoring. `gsem` still provides the theoretical ability to fit these models in multilevel contexts, but convergence may be difficult to achieve.

See the following example:

1. [SEM] **Example 17**. Correlated uniqueness model.

Latent growth models

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, we can write this as

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

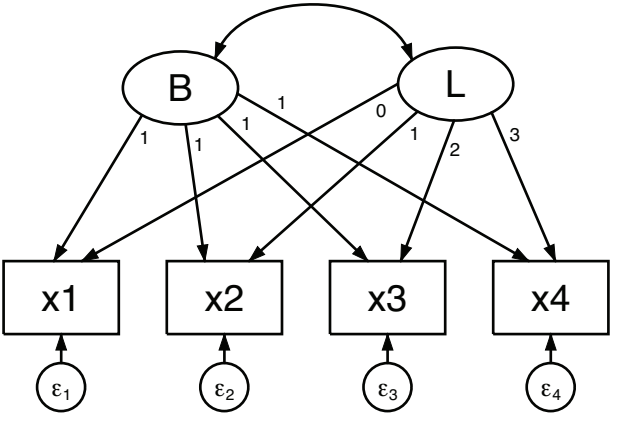
Let's assume that the observed values are collected over time. `x1` is observed at time 0, `x2` at time 1, and so on. It thus may be more reasonable to assume that the observed values represent a base value and a growth modeled with a linear trend. Thus we might write the model as

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

The equations for this model are

$$
\begin{aligned}
x_1 &= B + 0L + e.x_1 \\
x_2 &= B + 1L + e.x_2 \\
x_3 &= B + 2L + e.x_3 \\
x_4 &= B + 3L + e.x_4
\end{aligned}
$$

and the path diagram 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 Builder.

Latent growth models can be fit with sem or gsem.

See the following example:

1. [SEM] **Example 18**. Latent growth model.

Models with reliability

A typical solution for dealing with variables measured with error is to find multiple measurements and 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 and gsem option reliability( )**.

Models with reliability can be fit with sem or gsem, although option reliability() is available only when responses are Gaussian with the identity link and only in the absence of censoring.

See the following example:

1. [SEM] **Example 24**. Reliability.

Multilevel mixed-effects models

Multilevel modeling concerns the inclusion of common, random effects across groups of the data, which are known as levels. You have observations on students. Students attend schools. Or you have observations on patients. Patients are served by hospitals. In either case, there may be an unmeasured effect of the institution. Some schools are better than others while some schools are worse. The same applies to hospitals. These unmeasured effects can be parameterized as a latent variable that is constant within institution and varies across institution.

What we have just described is a two-level nested model. The first level, the lowest, is the observational level. The second level is school or hospital.

In a three-level nested model, students are served by schools which are served by counties, or patients are served by hospitals which are served by states. Whatever the example, there can be unmeasured effects of each of those higher levels contributing to the effect observed at the observational level.

An alternative to nested models is crossed models. People with jobs work in an industry and in a state. The same industries are found across states and the same states are found across industries, yet both may have an effect on some aspect of the lives of their workers.

Multilevel models are fit by gsem.

See the following examples:

1. [SEM] **Example 30g**. Two-level measurement model (multilevel, generalized response).
2. [SEM] **Example 38g**. Random-intercept and random-slope models (multilevel).
3. [SEM] **Example 39g**. Three-level model (multilevel, generalized response).
4. [SEM] **Example 40g**. Crossed models (multilevel).
5. [SEM] **Example 41g**. Two-level multinomial logistic regression (multilevel).
6. [SEM] **Example 42g**. One- and two-level mediation models (multilevel).

Latent class models

A latent class model involves a latent variable that is categorical rather than continuous. The unobserved levels of the categorical latent variable are called latent classes. These classes correspond to unobserved groups in the population such as unobserved groups of healthy and unhealthy individuals or unobserved groups of consumers with different buying preferences. Latent class analysis can help us to identify and understand these groups.

When we fit a latent class model, we specify the number of latent classes and then estimate the probabilities of class membership. In addition, the parameters that are estimated in the rest of the model are allowed to vary across the classes. For example, we may fit intercept-only logistic regression models to a set of binary variables and allow the intercepts to be estimated separately across classes.

Latent class models are fit by `gsem`. The Stata command syntax for this type of model is

```
(y1 y2 y3 y4 <- ), logit lclass(C 2)
```

This fits a latent class model with one categorical latent variable, `C`, that has two classes. Both the name of the latent variable and the number of classes is specified in the `lclass()` option.

This basic latent class model can be extended in many ways.

1. We can specify that `C` has three, four, or even more latent classes.
2. We are not limited to having observed variables that are categorical. When variables are continuous, the analysis is sometimes called by other names such as latent profile analysis or latent cluster analysis instead of latent class analysis.
3. We can include predictors of `C`—predictors of the probabilities of being in the different classes.
4. We can include more than one categorical latent variable.

See the following examples:

1. [SEM] **Example 50g**. Latent class model.
2. [SEM] **Example 51g**. Latent class goodness-of-fit statistics.
3. [SEM] **Example 52g**. Latent profile model.

Finite mixture models

Finite mixture models also involve categorical latent variables. Here we focus on finite mixture regression models in which you can fit any regression model allowed by gsem and estimate the parameters of that model separately for each latent class. For a linear regression of y on x1 and x2, the command syntax for a two-class model is

```
(y <- x1 x2), lclass(C 2)
```

The intercept and the coefficients on x1 and x2 will be estimated separately for the two classes. In addition, we estimate the probability of being in each class. If we have a variable z that predicts class membership, the command syntax becomes

```
(y <- x1 x2) (C <- z), lclass(C 2)
```

The fmm: prefix can also be used to fit these finite mixture regression models. For instance, fmm: regress can fit the model shown above. gsem extends the types of models that can be fit using fmm by allowing more than one response variable or more than one categorical latent variable.

See the following examples:

1. [SEM] **Example 53g**. Finite mixture Poisson regression.
2. [SEM] **Example 54g**. Finite mixture Poisson regression, multiple responses.

References

Acock, A. C. 2013. *Discovering Structural Equation Modeling Using Stata*. Rev. ed. College Station, TX: Stata Press.

Bartus, T. 2017. Multilevel multiprocess modeling with gsem. *Stata Journal* 17: 442–461.

Bauldry, S. 2014. miivfind: A command for identifying model-implied instrumental variables for structural equation models in Stata. *Stata Journal* 14: 60–75.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: Wiley. https://doi.org/10.1002/9781118619179.

Comulada, W. S. 2021. Calculating level-specific SEM fit indices for multilevel mediation analyses. *Stata Journal* 21: 195–205.

Palmer, T. M., and J. A. C. Sterne. 2015. Fitting fixed- and random-effects meta-analysis models using structural equation modeling with the sem and gsem commands. *Stata Journal* 15: 645–671.

Pickles, A., M. Bluett-Duncan, H. Sharp, and S. Vitoratou. 2024. Distinguishing differences in construct from differences in response style: gsem for item response theory models with anchoring vignettes. *Stata Journal* 24: 666–686.

Also see

[SEM] **Intro 4** — Substantive concepts

[SEM] **Intro 6** — Comparing groups

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