Example 1 — Single-factor measurement model

Description Remarks and examples References Also see

Description

The single-factor measurement model is demonstrated using the following data:

```
. use https://www.stata-press.com/data/r19/sem_1fmm
(Single-factor measurement model)
. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
x1	500	99.518	14.35402	60	137
x2	500	99.954	14.1939	52	140
x3	500	99.052	14.26395	59	150
x4	500	94.474	70.11603	-113	295

```
. notes
_dta:
  1.  fictional data
  2.  Variables x1, x2, x3, and x4 each contain a test score designed to
      measure X. The test is scored to have mean 100.
```

See *Single-factor measurement models* in [SEM] **Intro 5** for background.

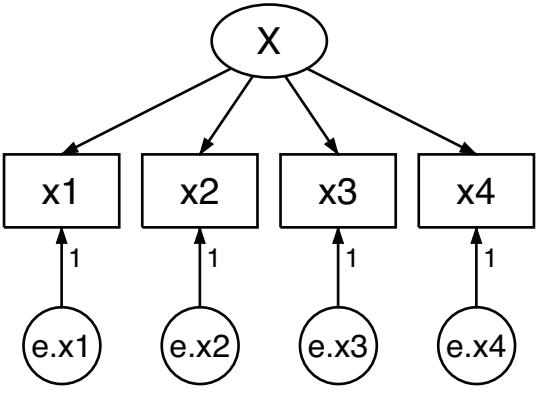
Remarks and examples

Remarks are presented under the following headings:

Single-factor measurement model
Satorra–Bentler scaled χ^2 test
Fitting the same model with gsem
Fitting the same model with the Builder
The measurement-error model interpretation

Single-factor measurement model

Below we fit the following model:



```
. sem (x1 x2 x3 x4 <- X)
Endogenous variables
  Measurement: x1 x2 x3 x4
Exogenous variables
  Latent: X
Fitting target model:
Iteration 0:  Log likelihood = -8487.5905
Iteration 1:  Log likelihood = -8487.2358
Iteration 2:  Log likelihood = -8487.2337
Iteration 3:  Log likelihood = -8487.2337
Structural equation model                                Number of obs = 500
Estimation method: ml
Log likelihood = -8487.2337
 ( 1)  [x1]X = 1
```

	Coefficient	OIM std. err.	z	P>|z|	[95% conf.	interval]
Measurement						
x1						
X	1	(constrained)				
_cons	99.518	.6412888	155.18	0.000	98.2611	100.7749
x2						
X	1.033249	.0723898	14.27	0.000	.8913676	1.17513
_cons	99.954	.6341354	157.62	0.000	98.71112	101.1969
x3						
X	1.063876	.0729725	14.58	0.000	.9208526	1.2069
_cons	99.052	.6372649	155.43	0.000	97.80298	100.301
x4						
X	7.276754	.4277638	17.01	0.000	6.438353	8.115156
_cons	94.474	3.132547	30.16	0.000	88.33432	100.6137
var(e.x1)	115.6865	7.790423			101.3823	132.0089
var(e.x2)	105.0445	7.38755			91.51873	120.5692
var(e.x3)	101.2572	7.17635			88.12499	116.3463
var(e.x4)	144.0406	145.2887			19.94838	1040.069
var(X)	89.93921	11.07933			70.64676	114.5001

```
LR test of model vs. saturated: chi2(2) = 1.46              Prob > chi2 = 0.4827
```

The equations for this model are

$$x_1 = \alpha_1 + X\beta_1 + e.x_1$$

$$x_2 = \alpha_2 + X\beta_2 + e.x_2$$

$$x_3 = \alpha_3 + X\beta_3 + e.x_3$$

$$x_4 = \alpha_4 + X\beta_4 + e.x_4$$

Notes:

1. Variable `X` is latent exogenous and thus needs a normalizing constraint. The variable is anchored to the first observed variable, `x1`, and thus the path coefficient is constrained to be 1. See *Identification 2: Normalization constraints (anchoring)* in [SEM] **Intro 4**.

2. The path coefficients for `X->x1`, `X->x2`, and `X->x3` are 1 (constrained), 1.03, and 1.06. Meanwhile, the path coefficient for `X->x4` is 7.28. This is not unexpected; we at StataCorp generated these data, and the true coefficients are 1, 1, 1, and 7.

3. A test for "model versus saturated" is reported at the bottom of the output; the $\chi^2(2)$ statistic is 1.46 and its significance level is 0.4827. We cannot reject the null hypothesis of this test.

   This test is a goodness-of-fit test in badness-of-fit units; a significant result implies that the model does not fit well.

   More mathematically, the null hypothesis of the test is that the fitted covariance matrix and mean vector of the observed variables are equal to the matrix and vector observed in the population.

Satorra–Bentler scaled χ^2 test

The model-versus-saturated goodness-of-fit statistic shown above does not follow the χ^2 distribution that it is referred to when the data are nonnormal. Satorra and Bentler (1994) provide a scaled version of this statistic that more closely follows the mean of the reference distribution in the presence of nonnormal data. We can request this statistic and the corresponding robust standard errors by specifying the `vce(sbentler)` option.

```
. sem (x1 x2 x3 x4 <- X), vce(sbentler)

Endogenous variables
  Measurement: x1 x2 x3 x4

Exogenous variables
  Latent: X

Fitting target model:
Iteration 0:  Log pseudolikelihood = -8487.5905
Iteration 1:  Log pseudolikelihood = -8487.2358
Iteration 2:  Log pseudolikelihood = -8487.2337
Iteration 3:  Log pseudolikelihood = -8487.2337

Structural equation model                                Number of obs = 500
Estimation method: ml

Log pseudolikelihood = -8487.2337

 ( 1)  [x1]X = 1
```

	Satorra-Bentler				
	Coefficient	std. err.	z	P>|z|	[95% conf. interval]
Measurement					
x1					
X	1	(constrained)			
_cons	99.518	.6419311	155.03	0.000	98.25984 100.7762
x2					
X	1.033249	.0767608	13.46	0.000	.8828006 1.183698
_cons	99.954	.6347705	157.46	0.000	98.70987 101.1981
x3					
X	1.063876	.0751028	14.17	0.000	.9166773 1.211075
_cons	99.052	.6379032	155.28	0.000	97.80173 100.3023
x4					
X	7.276754	.4386592	16.59	0.000	6.416998 8.13651
_cons	94.474	3.135684	30.13	0.000	88.32817 100.6198
var(e.x1)	115.6865	7.744173			101.4617 131.9055
var(e.x2)	105.0445	6.499187			93.04833 118.5872
var(e.x3)	101.2572	7.00047			88.42551 115.9509
var(e.x4)	144.0406	145.6607			19.84766 1045.347
var(X)	89.93921	11.2763			70.34416 114.9927

```
LR test of model vs. saturated: chi2(2) = 1.46           Prob > chi2 = 0.4827
Satorra-Bentler scaled test:    chi2(2) = 1.59           Prob > chi2 = 0.4526
```

The rescaled statistic is labeled “Satorra–Bentler scaled test” and has a value of 1.59 with a significance level of 0.4526. As with the unadjusted test, we cannot reject the null hypothesis.

Fitting the same model with gsem

sem and gsem produce the same results for standard linear SEMs. Let's demonstrate that.

```
. gsem (x1 x2 x3 x4 <- X)
Fitting fixed-effects model:
Iteration 0:  Log likelihood = -8948.2394
Iteration 1:  Log likelihood = -8948.2394
Refining starting values:
Grid node 0:  Log likelihood = -8487.5916
Fitting full model:
Iteration 0:  Log likelihood = -8487.5916
Iteration 1:  Log likelihood = -8487.5051
Iteration 2:  Log likelihood = -8487.3836
Iteration 3:  Log likelihood = -8487.2697
Iteration 4:  Log likelihood = -8487.2337
Iteration 5:  Log likelihood = -8487.2337
Generalized structural equation model                     Number of obs = 500
Response: x1
Family:   Gaussian
Link:     Identity
Response: x2
Family:   Gaussian
Link:     Identity
Response: x3
Family:   Gaussian
Link:     Identity
Response: x4
Family:   Gaussian
Link:     Identity
Log likelihood = -8487.2337
 ( 1)  [x1]X = 1
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
x1						
X	1	(constrained)				
_cons	99.518	.6412888	155.18	0.000	98.2611	100.7749
x2						
X	1.033249	.0723898	14.27	0.000	.8913676	1.17513
_cons	99.954	.6341354	157.62	0.000	98.71112	101.1969
x3						
X	1.063876	.0729725	14.58	0.000	.9208526	1.206899
_cons	99.052	.6372649	155.43	0.000	97.80298	100.301
x4						
X	7.276753	.4277636	17.01	0.000	6.438352	8.115154
_cons	94.474	3.132547	30.16	0.000	88.33432	100.6137
var(X)	89.93923	11.07933			70.64678	114.5001
var(e.x1)	115.6865	7.790422			101.3822	132.0089
var(e.x2)	105.0444	7.387549			91.51872	120.5692
var(e.x3)	101.2572	7.176349			88.12498	116.3463
var(e.x4)	144.0408	145.2886			19.94848	1040.067

Notes:

1. Results are virtually the same. Coefficients, variance estimates, and standard errors may differ in the last digit; for instance, `x4<-X` was 7.276754 and now it is 7.276753.

   These are the kind of differences we would expect to see. `gsem` follows a different approach for obtaining results that involves far more numeric machinery, which correspondingly results in slightly less accuracy.

2. The log-likelihood values reported are the same. This model is one of the few models we could have chosen where `sem` and `gsem` would produce the same log-likelihood values. In general, `gsem` log likelihoods are on different metrics from those of `sem`. In the case where the model does not include observed exogenous variables, however, they share the same metric.

3. There is no reason to use `gsem` over `sem` when both can fit the same model. `sem` is slightly more accurate, is quicker, and has more postestimation features.

Fitting the same model with the Builder

Use the diagram above for reference.

1. Open the dataset.

   In the Command window, type

   ```
   . use https://www.stata-press.com/data/r19/sem_1fmm
   ```

2. Open a new Builder diagram.

   Select menu item **Statistics > SEM (structural equation modeling) > Model building and estimation**.

3. Create the measurement component for `X`.

   Select the Add measurement component tool, , and then click in the diagram about one-third of the way down from the top and slightly left of the center.

   In the resulting dialog box,

   a. change the *Latent variable name* to `X`;

   b. select `x1`, `x2`, `x3`, and `x4` by using the *Measurement variables* control;

   c. select `Down` in the *Measurement direction* control;

   d. click on **OK**.

   If you wish, move the component by clicking on any variable and dragging it.

   Notice that the constraints of 1 on the paths from the error terms to the observed measures are implied, so we do not need to add these to our diagram.

4. Estimate.

   Click on the **Estimate** button, , in the Standard Toolbar, and then click on **OK** in the resulting *SEM estimation options* dialog box.

You can open a completed diagram in the Builder by typing

```
. webgetsem sem_1fmm
```

The measurement-error model interpretation

As we pointed out in *Using path diagrams to specify standard linear SEMs* in [SEM] **Intro 2**, if we rename variable x4 to be y, we can reinterpret this measurement model as a measurement-error model. In this interpretation, X is the unobserved true value. x1, x2, and x3 are each measurements of X, but with error. Meanwhile, y (x4) is really something else entirely. Perhaps y is earnings, and we believe

$$\texttt{y} = \alpha_4 + \beta_4\texttt{X} + \texttt{e.y}$$

We are interested in β_4, the effect of true X on y.

If we were to go back to the data and type `regress y x1`, we would obtain an estimate of β_4, but we would expect that estimate to be biased toward 0 because of the errors-in-variable problem. The same applies for y on x2 and y on x3. If we do that, we obtain

β_4 based on `regress y x1`	3.19
β_4 based on `regress y x2`	3.36
β_4 based on `regress y x3`	3.43

In the sem output above, we have an estimate of β_4 with the bias washed away:

β_4 based on `sem (y<-X)`	7.28

The number 7.28 is the value reported for `(x4<-X)` in the sem output.

That β_4 might be 7.28 seems plausible because we expect that the estimate should be larger than the estimates we obtain using the variables measured with error. In fact, we can tell you that the 7.28 estimate is quite good because we at StataCorp know that the true value of β_4 is 7. Here is how we manufactured this fictional dataset:

```
set seed 83216
set obs 500
gen X  = round(rnormal(0,10))
gen x1 = round(100 + X + rnormal(0, 10))
gen x2 = round(100 + X + rnormal(0, 10))
gen x3 = round(100 + X + rnormal(0, 10))
gen x4 = round(100 + 7*X + rnormal(0, 10))
drop X
```

The data recorded in `sem_1fmm.dta` were obviously generated using normality, the same assumption that is most often used to justify the SEM maximum likelihood estimator. In [SEM] **Intro 4**, we explained that the normality assumption can be relaxed and conditional normality can usually be substituted in its place.

So let's consider nonnormal data. Let's make X be $\chi^2(2)$, a violently nonnormal distribution, resulting in the data-manufacturing code

```
set seed 83216
set obs 500
gen X  = (rchi2(2)-2)*(10/2)
gen x1 = round(100 + X + rnormal(0, 10))
gen x2 = round(100 + X + rnormal(0, 10))
gen x3 = round(100 + X + rnormal(0, 10))
gen x4 = round(100 + 7*X + rnormal(0, 10))
drop X
```

All the `rnormal()` functions remaining in our code have to do with the assumed normality of the errors. The multiplicative and additive constants in the generation of `X` simply rescale the $\chi^2(2)$ variable to have mean 100 and standard deviation 10, which would not be important except for the subsequent `round()` functions, which themselves were unnecessary except that we wanted to produce a pretty dataset when we created the original `sem_1fmm.dta`.

In any case, if we rerun the commands with these data, we obtain

β_4 based on `regress y x1`	3.24
β_4 based on `regress y x2`	3.14
β_4 based on `regress y x3`	3.36
β_4 based on `sem (y<-X)`	7.25

We will not burden you with the details of running simulations to assess coverage; we will just tell you that coverage is excellent: reported test statistics and significance levels can be trusted.

By the way, errors in the variables is something that does not go away with progressively larger sample sizes. Change the code above to produce a 100,000-observation dataset instead of a 500-observation one, and you will obtain

β_4 based on `regress y x1`	3.47
β_4 based on `regress y x2`	3.48
β_4 based on `regress y x3`	3.50
β_4 based on `sem (y<-X)`	6.97

References

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

Satorra, A., and P. M. Bentler. 1994. "Corrections to test statistics and standard errors in covariance structure analysis". In *Latent Variables Analysis: Applications for Developmental Research*, edited by A. von Eye and C. C. Clogg, 399–419. Thousand Oaks, CA: Sage.

Also see

[SEM] **sem** — Structural equation model estimation command

[SEM] **gsem** — Generalized structural equation model estimation command

[SEM] **Intro 5** — Tour of models

[SEM] **Example 3** — Two-factor measurement model

[SEM] **Example 24** — Reliability

[SEM] **Example 27g** — Single-factor measurement model (generalized response)

Stata, Stata Press, Mata, NetCourse, and NetCourseNow are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow is a trademark of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.

For suggested citations, see the FAQ on citing Stata documentation.