

## example 4 — Goodness-of-fit statistics

[Description](#)[Remarks and examples](#)[Reference](#)[Also see](#)

## Description

Here we demonstrate `estat gof`. See [\[SEM\] intro 7](#) and [\[SEM\] estat gof](#).

This example picks up where [\[SEM\] example 3](#) left off:

```
. use http://www.stata-press.com/data/r15/sem_2fmm
. sem (Affective -> a1 a2 a3 a4 a5) (Cognitive -> c1 c2 c3 c4 c5)
```

## Remarks and examples

stata.com

When we fit this model in [\[SEM\] example 3](#), at the bottom of the output, we saw

```
. sem (Affective -> a1 a2 a3 a4 a5) (Cognitive -> c1 c2 c3 c4 c5)
(output omitted)
LR test of model vs. saturated: chi2(34) = 88.88, Prob > chi2 = 0.0000
```

Most texts refer to this test against the saturated model as the “model  $\chi^2$  test”.

These results indicate poor goodness of fit; see [\[SEM\] example 1](#). The default goodness-of-fit statistic reported by `sem`, however, can be overly influenced by sample size, correlations, variance unrelated to the model, and multivariate nonnormality ([Kline 2016, 271](#)).

Goodness of fit in cases of `sem` is a measure of how well you fit the observed moments, which in this case are the covariances between all pairs of `a1, …, a5, c1, …, c5`. In a measurement model, the assumed underlying causes are unobserved, and in this example, those unobserved causes are the latent variables `Affective` and `Cognitive`. It may be reasonable to assume that the observed `a1, …, a5, c1, …, c5` can be filtered through imagined variables `Affective` and `Cognitive`, but that can be reasonable only if not too much information contained in the original variables is lost. Thus goodness-of-fit statistics are of great interest to those fitting measurement models. Goodness-of-fit statistics are of far less interest when all variables in the model are observed.

Other goodness-of-fit statistics are available.

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

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(34)	88.879	model vs. saturated
p > chi2	0.000	
chi2_bs(45)	2467.161	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.086	Root mean squared error of approximation
90% CI, lower bound	0.065	
upper bound	0.109	
pclose	0.004	Probability RMSEA <= 0.05
Information criteria		
AIC	19120.770	Akaike's information criterion
BIC	19191.651	Bayesian information criterion
Baseline comparison		
CFI	0.977	Comparative fit index
TLI	0.970	Tucker-Lewis index
Size of residuals		
SRMR	0.022	Standardized root mean squared residual
CD	0.995	Coefficient of determination

Notes:

1. Desirable values vary from test to test.
2. We asked for all the goodness-of-fit tests. We could have obtained specific tests from the above output by specifying the appropriate option; see [\[SEM\] estat gof](#).
3. Under likelihood ratio, `estat gof` reports two tests. The first is a repeat of the model  $\chi^2$  test reported at the bottom of the `sem` output. The saturated model is the model that fits the covariances perfectly. We can reject at the 5% level (or any other level) that the model fits as well as the saturated model.

The second test is a baseline versus saturated comparison. The baseline model includes the mean and variances of all observed variables plus the covariances of all observed exogenous variables. Different authors define the baseline differently. We can reject at the 5% level (or any other level) that the baseline model fits as well as the saturated model.

4. Under population error, the RMSEA value is reported along with the lower and upper bounds of its 90% confidence interval. Most interpreters of this test check whether the lower bound is below 0.05 or the upper bound is above 0.10. If the lower bound is below 0.05, then they would not reject the hypothesis that the fit is close. If the upper bound is above 0.10, they would not reject the hypothesis that the fit is poor. The logic is to perform one test on each end of the 90% confidence interval and thus have 95% confidence in the result. This model's fit is not close, and its upper limit is just over the bounds of being considered poor.

Pclose, a commonly used word in reference to this test, is the probability that the RMSEA value is less than 0.05, interpreted as the probability that the predicted moments are close to the moments in the population. This model's fit is not close.

5. Under information criteria are reported AIC and BIC, which contain little information by themselves but are often used to compare models. Smaller values are considered better.
6. Under baseline comparison are reported CFI and TLI, two indices such that a value close to 1 indicates a good fit. TLI is also known as the nonnormed fit index.
7. Under size of residuals is reported the standardized root mean squared residual (SRMR) and the coefficient of determination (CD).

A perfect fit corresponds to an SRMR of 0, and a good fit corresponds to a “small” value, considered by some to be limited at 0.08. The model fits well by this standard.

The CD is like an  $R^2$  for the whole model. A value close to 1 indicates a good fit.

`estat gof` provides multiple goodness-of-fit statistics because, across fields, different researchers use different statistics. You should not print them all and look for the one reporting the result you seek.

## Reference

Kline, R. B. 2016. *Principles and Practice of Structural Equation Modeling*. 4th ed. New York: Guilford Press.

## Also see

[SEM] [example 3](#) — Two-factor measurement model

[SEM] [example 21](#) — Group-level goodness of fit

[SEM] [estat gof](#) — Goodness-of-fit statistics