“What are the saturated and baseline models in SEM?”

Authored by

stats writer

July 1, 2024

SEM (structural equation modeling) is a statistical method used to analyze complex relationships between variables. In SEM, there are two main types of models: saturated and baseline models.

A saturated model is a theoretical model that perfectly fits the observed data. This means that all of the variables in the model are directly connected to one another, resulting in a perfect fit. Saturated models are used as a benchmark to compare other models against, as they represent the best possible fit to the data.

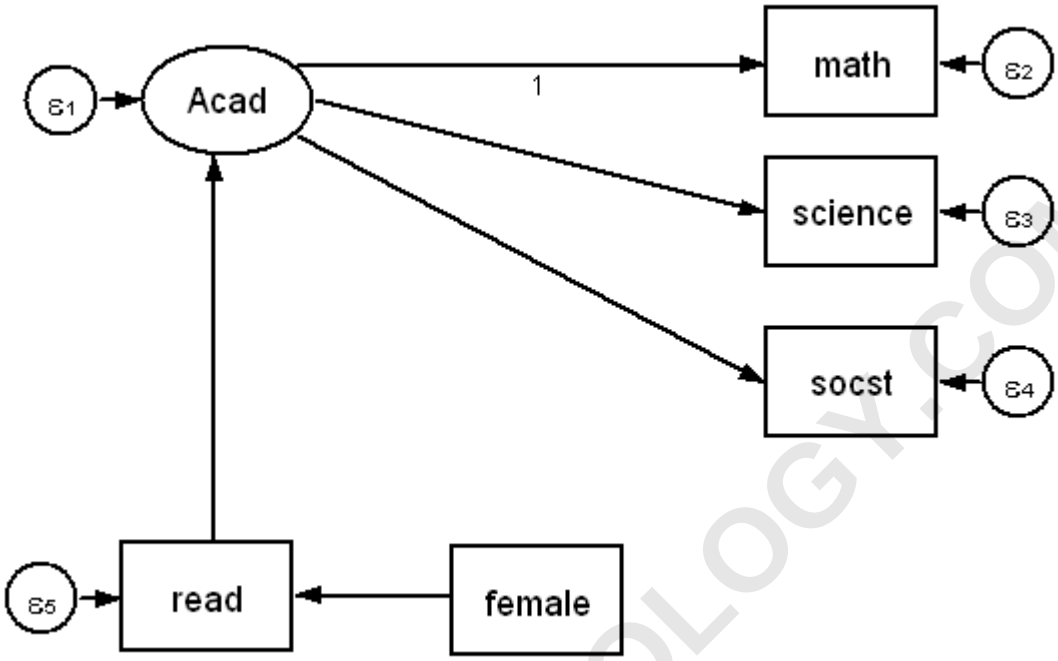
On the other hand, a baseline model is a simplified version of the saturated model that assumes no relationships between variables. This model is also known as the null model, as it serves as a starting point for evaluating the fit of more complex models. The baseline model provides a point of comparison for determining if a more complex model is necessary to explain the relationships between variables.

In summary, the saturated model represents the ideal fit for the data, while the baseline model serves as a starting point for evaluating the fit of more complex models in SEM. Both models play an important role in the evaluation and interpretation of relationships between variables in SEM.

What are the saturated and baseline models in sem? | Stata FAQ

Below is the diagram of a simple structural equation model. The dependent variable is a latent variable Acad with three observed indicators, math, science and socst. There are two additional observed variables, the independent variable female and a mediator variable read. (Note, variables in squares are observed (manifest variables), those in circles are latent. The small circles with ε are error terms, i.e., residual

variances).



We will analyze this model using the sem command with the hsbdemo dataset.

use https://stats.idre.ucla.edu/stat/data/hsbdemo, clear

sem (Acad->math science socst)(Acad

Endogenous variables

Observed: read
Measurement: math science socst
Latent: Acad

Exogenous variables

Observed: female

Fitting target model:

Iteration 0: log likelihood = -6737.783 (not concave)

Iteration 13: log likelihood = -2949.3343

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -2949.3343

(1) Acad = 1

--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Structural |
read chi2 = 0.0315

estat gof

--
Fit statistic | Value Description

----------+----------

Likelihood ratio |

chi2_ms(5) | 12.251 model vs. saturated

p > chi2 | 0.032

chi2_bs(10) | 361.012 baseline vs. saturated

p > chi2 | 0.000

The estat gof makes reference to three different models; 1) the model (the one we just ran), 2) the saturated model, and 3) the baseline model. Before we discuss the saturated and baseline models, let's look a little closer at the above model.

In the above model we estimated 15 parameters; 2 structural coefficients, 1 structural intercept, 2 measurement coefficients (loadings), 3 measurement intercepts, 6 variances and 1 mean. The log likelihood for our model was -2949.3343.

The saturated model

Now let's move on to the saturated model. A saturated model perfectly reproduces all

of the variances, covariance and means of the observed variables. Here is a simple way to produce a saturated model.

```
sem (

Exogenous variables

Observed: read math science socst female

Fitting target model:

Iteration 0: log likelihood = -2943.2087
Iteration 1: log likelihood = -2943.2087

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -2943.2087

------------------------------------------------------------------------------
| OIM
| Coef. Std. Err. z P>|z|
-------------+----------------------------------------------------------------
Mean |
read | 52.23 .7231774 72.22 0.000 50.8126 53.6474
math | 52.645 .6607911 79.67 0.000 51.34987 53.94013
```

science | 51.85 .6983463 74.25 0.000 50.48127 53.21873
socst | 52.405 .757235 69.21 0.000 50.92085 53.88915
female | .545 .0352119 15.48 0.000 .475986 .614014
-------------+--
Variance |
read | 104.5971 10.45971 85.98041 127.2447
math | 87.32898 8.732897 71.78574 106.2377
science | 97.5375 9.75375 80.17731 118.6566
socst | 114.681 11.4681 94.2695 139.512
female | .247975 .0247975 .2038392 .3016672
-------------+--
Covariance |
read |
math | 63.29665 8.105808 7.81 0.000 47.40956 79.18374
science | 63.6495 8.441978 7.54 0.000 47.10353 80.19547
socst | 68.06685 9.118222 7.46 0.000 50.19546 85.93824
female | -.27035 .3606283 -0.75 0.453 -.9771685 .4364685
-----------+--
math |
science | 58.21175 7.715717 7.54 0.000 43.08922 73.33428
socst | 54.48877 8.057294 6.76 0.000 38.69677 70.28078
female | -.136525 .3291963 -0.41 0.678 -.7817379 .5086879

```
-----------+------------------------------------------------------------
science |
socst | 49.19075 8.247856 5.96 0.000 33.02525 65.35625
female | -.62825 .3505821 -1.79 0.073 -1.315378 .0588783
-----------+------------------------------------------------------------
socst |
female | .279275 .3775977 0.74 0.460 -.460803 1.019353
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0) = 0.00, Prob > chi2 = .
```

A saturated model has the best fit possible since it perfectly reproduces all of the
variances, covariances and means. That's why the saturated model above has a chi-square of
zero with zero degrees of freedom. Since you can't do any better than a saturated
model, it becomes the standard for comparison with the models that you estimate.

For the saturated model we estimated 20 parameters; 5 variances, 10 covariances and
5 means. You can compute the number of parameters in a saturated model of k

observed variables by the formula k*(k+1)/2 + k. In our example, it is
5*(5+1)/2 + 5 = 20. The log likelihood for this model is -2943.2087.

To test how well our model compares to a saturated model, we compute chi-square as follows,
minus two times the differences in the log likelihoods; -2*(-2949.3343 - -2943.2087) = 12.2512.
The degrees of freedom for this chi-square is the difference in the number of parameters estimated
in the two model (20 - 15 = 5). Thus, our model fits significantly poorer than a saturated
model (p = .0315). But, that's not surprising since our model was only for demonstration
purposes.

The baseline model

So, that brings us to the baseline model. This is defined in the Stata Structural
Equation Modeling Reference Manual as a model which includes the means and variances
of all observed variables plus the covariances of all observed exogenous variables. Since there

is only one observed exogenous variable, female, in our model, there will be no
covariances in our baseline model.

sem (

Exogenous variables

Observed: read math science socst female

Fitting target model:

Iteration 0: log likelihood = -3123.7147
Iteration 1: log likelihood = -3123.7147

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -3123.7147

(1) _cons = 0
(2) _cons = 0
(3) _cons = 0
(4) _cons = 0
(5) _cons = 0
(6) _cons = 0
(7) _cons = 0

(8) _cons = 0

(9) _cons = 0

(10) _cons = 0

| OIM

| Coef. Std. Err. z P>|z|

-------------+--

Mean |

read | 52.23 .7231774 72.22 0.000 50.8126 53.6474

math | 52.645 .6607911 79.67 0.000 51.34987 53.94013

science | 51.85 .6983463 74.25 0.000 50.48127 53.21873

socst | 52.405 .757235 69.21 0.000 50.92085 53.88915

female | .545 .0352119 15.48 0.000 .475986 .614014

-------------+--

Variance |

read | 104.5971 10.45971 85.98041 127.2447

math | 87.32898 8.732898 71.78574 106.2377

science | 97.5375 9.75375 80.17731 118.6566

socst | 114.681 11.4681 94.2695 139.512

female | .247975 .0247975 .2038392 .3016672

-------------+--

Covariance |

read |

math | 0 (constrained)

science | 0 (constrained)
socst | 0 (constrained)
female | 0 (constrained)
----------+----------
math |
science | 0 (constrained)
socst | 0 (constrained)
female | 0 (constrained)
----------+----------
science |
socst | 0 (constrained)
female | 0 (constrained)
----------+----------
socst |
female | 0 (constrained)

LR test of model vs. saturated: chi2(10) = 361.01, Prob > chi2 = 0.0000

For the baseline model we estimated 10 parameters; 5 variances and 5 means. In comparing
this model with the saturated model there was a difference of 10 degrees of freedom,
20 - 10 = 10. Again, we compute chi-square as minus

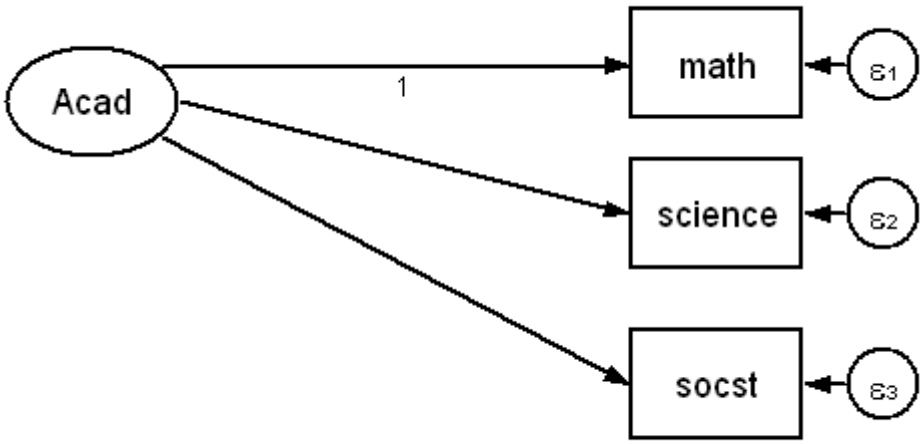
two times the difference in
the log likelihoods, -2*(-3123.7147 - -2943.2087) = 361.012.

Although our model did not fit all that well compared to the saturated model, the fit of
the baseline model compared to the saturated model is much worse, with chi2(10) =
361.012, p = 0.0000.

The two chi-square values from the estat gof for our model versus a saturated model
and baseline versus saturated model help us to understand how well our model fits the data.

The saturated model revisited

When we looked at the saturated model above we used a very simple model with only
observed variables. Now we are going to try to come up with a saturated model that
is more closely related to our original model. We will begin by looking at just the
measurement part of our model. Here is the diagram.



Followed by the sem code.

sem (Acad->math science socst)

Endogenous variables

Measurement: math science socst

Exogenous variables

Latent: Acad

Fitting target model:

Iteration 0: log likelihood = -2141.1294

Iteration 1: log likelihood = -2141.1294

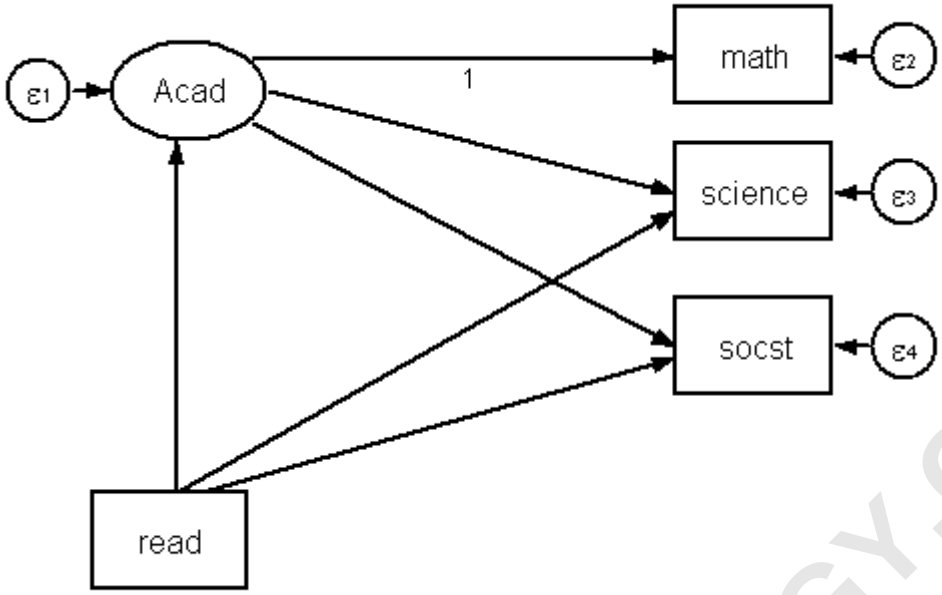
Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -2141.1294

(1) Acad = 1

--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Measurement |
math chi2 = .

As you can see, the measure model with three indicators is itself a saturated model. To be saturated it should have 3*4/2 + 3 = 9 parameters being estimated, which is the case.

Now, let's add read to our model like this.



sem (Acad->math science socst)(Acad

Endogenous variables

Observed: science socst

Measurement: math

Latent: Acad

Exogenous variables

Observed: read

Fitting target model:

Iteration 0: log likelihood = -3698.205 (not concave)

Iteration 28: log likelihood = -2802.3352

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -2802.3352

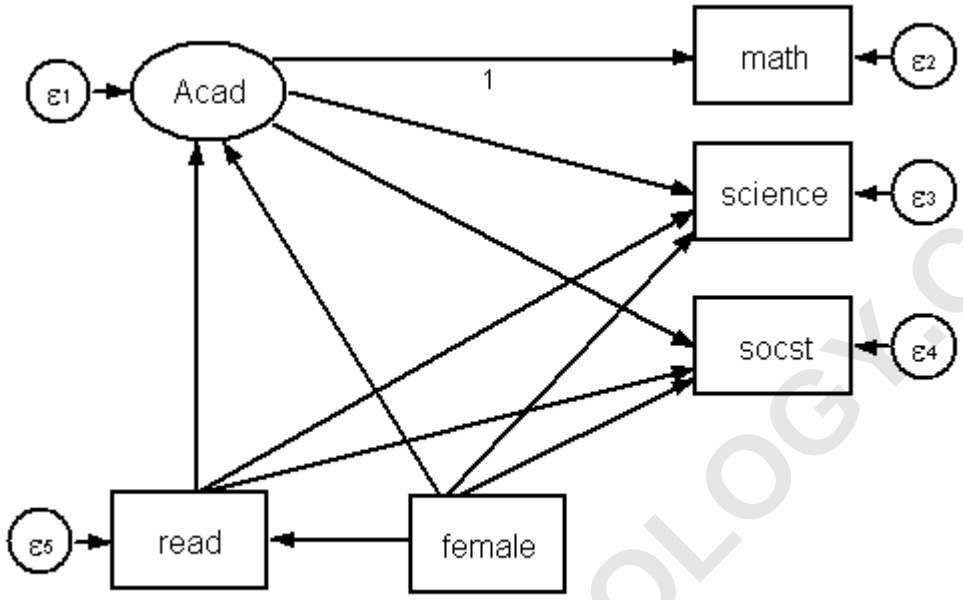
(1) Acad = 1

--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Structural |
science chi2 = .

This model has four observed variables. Thus, we should estimate 4*5/2 + 4 = 14
parameters. We achieved this by adding direct paths from read to science
and to socst. We could have also achieved the same result by adding two
covariances, say e.math*e.science and e.math*e.socst, to our model instead of
the direct effects.

Finally, let's add female to our model. We now have as

many observed variables as
our original model.



The above diagram translates to the following code.

sem (Acad -> math science socst)(Acad

Endogenous variables

Observed: math socst read
Measurement: science
Latent: Acad

Exogenous variables

Observed: female

Fitting target model:

Iteration 0: log likelihood = -3111.6647 (not concave)

Iteration 58: log likelihood = -2943.2087

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -2943.2087

(1) Acad = 1

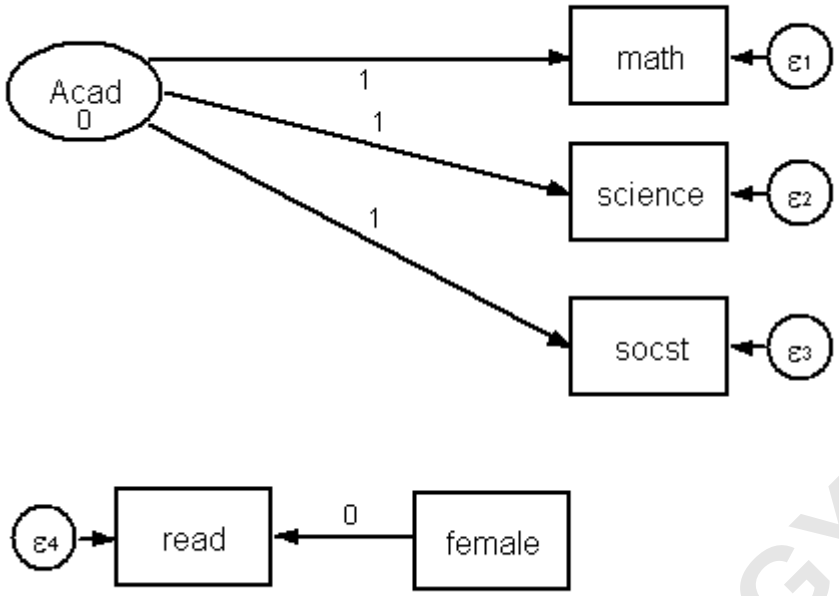
--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Structural |
math chi2 = .

This time there are five observed variables which means that we need to estimate 5*6/2 + 5 = 20 parameters for a saturated model. We did this by adding direct paths from female to
Acad, math, and socst and direct paths from read to

math
and socst. This is the same result that was obtained with the simpler approach
used earlier for the saturated model.

The baseline model revisited

We know that the baseline model estimates five means and five variances and no covariances,
because there is only one observed exogenous variables, for a total of 10 total
parameters. We can get this from our original model by constraining all of the
measurement coefficients (loadings) to be one and all of the path coefficients to be
zero. Here is a diagram of the model.



And, here is one way to accomplish this.

sem (Acad ->math@1 science@1 socst@1) (read

Endogenous variables

Observed: read
Measurement: math science socst

Exogenous variables

Observed: female
Latent: Acad

Fitting target model:

Iteration 0: log likelihood = -3257.7854

Iteration 4: log likelihood = -3123.7147

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -3123.7147

(1) Acad = 1
(2) Acad = 1
(3) Acad = 1
(4) _cons = 0

--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Structural |
read chi2 = 0.0000

In this model the term (read estimates an intercept (mean) but no
structural coefficient. There is no term that predicting Acad from read
which is equivalent to setting that structural coefficient to zero. We added terms

for the mean and variance of female. Finally, by convention, the variance
of the latent variables is constrained to zero, which we did.