How can I perform Confirmatory Factor Analysis (CFA) using binary variables in Stata?

Authored by

stats writer

July 1, 2024

Confirmatory Factor Analysis (CFA) is a statistical technique used to assess the factor structure of a set of variables. This method is commonly used in social sciences, psychology, and other fields to test the validity of a theoretical construct or measure. In Stata, CFA can be performed using binary variables, which are categorical variables with only two possible outcomes. To do so, the user must first specify the model, which includes the number of factors, the relationship between the factors and the binary variables, and any other relevant parameters. Then, the data can be analyzed using a specialized command, such as "cfa" or "sem". Stata provides various options and tools to aid in the interpretation and evaluation of the CFA results, such as goodness-of-fit indices and modification indices. Overall, Stata offers a user-friendly and efficient approach to performing CFA with binary variables, making it a valuable tool for researchers and practitioners in a variety of fields.

How can I do CFA with binary variables? | Stata FAQ

Let's say that you have a dataset with a bunch of binary variables. Further, you believe that these binary variables reflect underlying and unobserved continuous variables. You don't want to compute your confirmatory factor analysis (CFA) directly on the binary variables. You will want to compute the CFA on tetrachoric correlations that reflect the associations among these underlying continuous variables. We will demonstrate this by using data with five continuous variables and creating binary variables from them by dichotomizing them at a point a little above their mean values.

**Let's begin by loading the hsbdemo.dta dataset and creating binary variables for
read, write, math, science and socst.**

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

gen r=read
```

**Now that we have the binary variables, let's checkout the correlations among the continuous
version of the variables and the binary version.**

```
corr read write math science socst

(obs=200)

| read write math science socst
-------------+---------------------------------------------
read | 1.0000
write | 0.5968 1.0000
math | 0.6623 0.6174 1.0000
science | 0.6302 0.5704 0.6307 1.0000
socst | 0.6215 0.6048 0.5445 0.4651 1.0000

corr r w m s o
```

(obs=200)

| r w m s o
----------+---
r | 1.0000
w | 0.4109 1.0000
m | 0.4750 0.5029 1.0000
s | 0.3846 0.4320 0.4750 1.0000
o | 0.4057 0.3910 0.3676 0.3009 1.0000

As you can see, the correlations among the binary version of the variables are much lower than among the continuous version. The Pearson correlations tend to underestimate the relationship between the underlying continuous variables that give rise to the binary variables. What we need are the tetrachoric correlations which we can obtain using the tetrachoric command.

tetrachoric r w m s o

(obs=200)

| r w m s o

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

r | 1.0000

w | 0.6145 1.0000

m | 0.6874 0.7176 1.0000

s | 0.5790 0.6411 0.6874 1.0000

o | 0.6148 0.5780 0.5556 0.4690 1.0000

The tetrachoric correlations are much closer to the original correlations among the continuous variables than the correlations among the binary values.

For comparison purposes we will compute a CFA on the original continuous data.

sem (FC->read write math science socst)

Endogenous variables

Measurement: read write math science socst

Exogenous variables

Latent: FC

Fitting target model:

Iteration 0: log likelihood = -3469.2622

Iteration 3: log likelihood = -3468.8093

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -3468.8093

(1) FC = 1

--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Measurement |
read chi2 = 0.0163sem (FB->r w m s o)

Next, we will create the SSD dataset and compute the CFA on the tetrachoric correlations.

clear

ssd init r w m s o

Summary statistics data initialized. Next use, in any order,

ssd set observations (required)
It is best to do this first.

ssd set means (optional)
Default setting is 0.

ssd set variances or ssd set sd (optional)
Use this only if you have set or will set correlations and, even then,
this is optional but highly recommended. Default setting is 1.

ssd set covariances or ssd set correlations (required)

ssd set obs 200
(value set)

Status:
observations: set
means: unset
variances or sd: unset
covariances or correlations: unset (required to be set)

ssd set cor 1.0000 ///
0.6145 1.0000 ///
0.6874 0.7176 1.0000 ///

0.5790 0.6411 0.6874 1.0000 ///
0.6148 0.5780 0.5556 0.4690 1.0000
(values set)

Status:
observations: set
means: unset
variances or sd: unset
covariances or correlations: set

sem (FT->r w m s o)

Endogenous variables

Measurement: r w m s o

Exogenous variables

Latent: FT

Fitting target model:

Iteration 0: log likelihood = -1148.1182
Iteration 1: log likelihood = -1147.4763
Iteration 2: log likelihood = -1147.4673
Iteration 3: log likelihood = -1147.4673

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -1147.4673

(1) FT = 1

--
| OIM
| Coef. Std. Err. z P>|z|
-------------+--
Measurement |
r chi2 = 0.0124

You will note that the model fit versus a saturated model is very close to the value that
was obtained when ran the CFA on the continuous variables.