How can I do mediation analysis with the sem command?

Authored by

stats writer

July 1, 2024

Mediation analysis is a statistical method used to explore the relationship between an independent variable and a dependent variable through the inclusion of a mediator variable. One way to conduct mediation analysis is by using the sem command in statistical software, such as SPSS or Stata. This command allows for the estimation of path coefficients and indirect effects, as well as significance testing of the mediation effect. It also allows for the examination of multiple mediator variables and the inclusion of control variables. By using the sem command, researchers can gain a deeper understanding of the underlying mechanisms and processes involved in their research relationship.

How can I do mediation analysis with the sem command? | Stata FAQ

The sem command introduced in Stata 12 makes the analysis of mediation models much easier as long as both the dependent variable and the mediator variable are continuous variables.

We will illustrate using the sem command with the hsbdemo dataset. The examples will not demonstrate full mediation, i.e., the effect of the independent variable will not go from being significant to being not significant. Rather, the examples will show partial mediation in which there is a decrease in the direct effect.

A note about covariates

If your model contains control variables, i.e., covariates, you must include these in each of the sem equations. Thus, your sem model will look something like this:

sem (MV <- IV CV1 CV2)(DV <- MV IV CV1 CV2)

where DV stands for the dependent variable, IV stands for the independent variable, MV stands for the mediator variable, and CVs stand for the covariates.

Simple mediation model

The simplest mediation model had one IV, one MV and a DV. Here is the symbolic version of the model.

sem (MV <- IV)(DV <- MV IV)

In our simple mediation example the independent variable is math, the mediator variable is read and the dependent variable is

science.

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

sem (read <- math)(science <- read math)

Endogenous variables

Observed: read science

Exogenous variables

Observed: math

Fitting target model:

Iteration 0: log likelihood = -2098.5822
Iteration 1: log likelihood = -2098.5822

Structural equation model Number of obs = 200
Estimation method = ml
Log likelihood = -2098.5822

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

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

Structural |

read |

math | .724807 .0579824 12.50 0.000 .6111636 .8384504

_cons | 14.07254 3.100201 4.54 0.000 7.996255 20.14882

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

science |

read | .3654205 .0658305 5.55 0.000 .2363951 .4944459

math | .4017207 .0720457 5.58 0.000 .2605138 .5429276

_cons | 11.6155 3.031268 3.83 0.000 5.674324 17.55668

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

var(e.read)| 58.71925 5.871925 48.26811 71.43329

var(e.science)| 50.8938 5.08938 41.83548 61.91346

LR test of model vs. saturated: chi2(0) = 0.00, Prob > chi2 = .

estat teffects

Direct effects

| OIM

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

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

Structural |

read |

math | .724807 .0579824 12.50 0.000 .6111636 .8384504

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

science |

read | .3654205 .0658305 5.55 0.000 .2363951 .4944459

math | .4017207 .0720457 5.58 0.000 .2605138 .5429276

Indirect effects

| OIM

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

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

Structural |

read |

math | 0 (no path)

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

science |

read | 0 (no path)

math | .2648593 .0522072 5.07 0.000 .1625351 .3671836

Total effects

| OIM

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

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

Structural |

read |

math | .724807 .0579824 12.50 0.000 .6111636 .8384504

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

science |

read | .3654205 .0658305 5.55 0.000 .2363951 .4944459

math | .66658 .05799 11.49 0.000 .5529217 .7802384

--

The total effect for math, .66658, is the effect we would find if there was no mediator in our model. It is significant with a z of 11.49. The direct effect for math is .4017207 which, while still significant (z = 5.58), is much smaller than the total effect. The indirect effect of math that passes through read is .2648593 and is also statistically significant.

It is often easier to interpret these values by computing ratios and proportions as

shown below.

proportion of total effect mediated = .2648593/.66658 = .3973406

ratio of indirect to direct effect = .2648593/.4017207 = .65931205

ratio of total to direct effect = .66658/.4017207 = 1.6593121

We see above that the proportion of the total effect that is mediated is almost .40 which is a respectable amount. The ratio of the indirect effect to the direct effect is about .66 or almost 2/3 the size of the direct effect. And finally, the total effect is about 1.66 times the direct effect.

Mediation with bootstrap standard errors and confidence intervals

If you are uncomfortable with the standard errors and confidence intervals produced directly by sem, you can obtain the bootstrapped standard errors and confidence intervals in two ways. First, by using the `vce(boostrap)` option after your sem

command. Or second, by writing a small program that runs both the sem command and the `estat teffects` and then bootstrapping this program.

Let's demonstrate the vce(boostrap) option. Here we will add the reps option and request 200 replications.

sem (read <- math)(science <- read math), vce(bootstrap,reps(200))
Bootstrap replications (200)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.... 50
.... 100
.... 150
.... 200

Structural equation model Number of obs = 200
Estimation method = ml Replications = 200
Log likelihood = -2098.5822

| Observed Bootstrap Normal-based
| Coef. Std. Err. z P>|z|
-------------+--
Structural |

read |
math | .724807 .0581262 12.47 0.000 .6108818 .8387321
_cons | 14.07254 3.092117 4.55 0.000 8.012099 20.13297
------------+--
science |
read | .3654205 .0802203 4.56 0.000 .2081915 .5226495
math | .4017207 .0875101 4.59 0.000 .2302041 .5732373
_cons | 11.6155 2.707368 4.29 0.000 6.309158 16.92184
-------------+--
var(e.read)| 58.71925 5.93704 48.16332 71.58871
var(e.science)| 50.8938 5.496477 41.18471 62.89176

Adding this option provides us bootstrapped confidence intervals. You can now use estat teffects to obtain normal-based bootstrapped confidence intervals around the indirect effect.

estat teffects

Direct effects

| Observed Bootstrap Normal-based
| Coef. Std. Err. z P>|z|

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

Structural |

read |

math | .724807 .0581262 12.47 0.000 .6108818 .8387321

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

science |

read | .3654205 .0802203 4.56 0.000 .2081915 .5226495

math | .4017207 .0875101 4.59 0.000 .2302041 .5732373

Indirect effects

| Observed Bootstrap Normal-based

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

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

Structural |

read |

math | 0 (no path)

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

science |

read | 0 (no path)

math | .2648593 .0593311 4.46 0.000 .1485726 .3811461

Total effects

```
------------------------------------------------------------------------------
| Observed Bootstrap Normal-based
| Coef. Std. Err. z P>|z|
-------------+----------------------------------------------------------------
Structural |
read |
math | .724807 .0581262 12.47 0.000 .6108818 .8387321
-----------+----------------------------------------------------------------
science |
read | .3654205 .0802203 4.56 0.000 .2081915 .5226495
math | .66658 .0592669 11.25 0.000 .5504189 .7827411
------------------------------------------------------------------------------
```

However, you can also write a program to perform the bootstrapping. This enables us to obtain both percentile-based and bias-corrected confidence intervals as well as normal-based confidence intervals. Here is the program that we a calling indireff.ado.

```
program indireff, rclass
sem (read <- math)(science <- read math)
estat teffects
mat bi = r(indirect)
```

```
mat bd = r(direct)
mat bt = r(total)
return scalar indir = el(bi,1,3)
return scalar direct = el(bd,1,3)
return scalar total = el(bt,1,3)
end
```

So how do we know which elements of r(indirect), r(direct) and r(total)
we need? We will use the sem command and then quietly run estat teffects
followed by a matrix list to see the matrices of the coefficients.

```
sem (read <- math)(science <- read math)
quietly estat teffects

matrix list r(indirect)

r(indirect)
read: science: science:
o. o.
math read math
r1 0 0 .26485934
```

matrix list r(direct)

r(direct)
read: science: science:
math read math
r1 .72480697 .36542052 .40172068

matrix list r(total)

r(total)
read: science: science:
math read math
r1 .72480697 .36542052 .66658002

We see that in each case the coefficient of interest is the third element.

Now that we know the correct matrix elements, we will run indireff for 200 bootstrap replications.
You may want to run more, say 2,000 to 5,000. We will then request the percentile and
biased corrected confidence intervals.

set seed 358395

bootstrap r(indir) r(direct) r(total), reps(200): indireff

Bootstrap replications (200)

----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5

.. 50

.. 100

.. 150

.. 200

Bootstrap results Number of obs = 200

Replications = 200

command: indireff

_bs_1: r(indir)

_bs_2: r(direct)

_bs_3: r(total)

--

| Observed Bootstrap Normal-based

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

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

_bs_1 | .2648593 .0545941 4.85 0.000 .1578569 .3718618

_bs_2 | .4017207 .0872965 4.60 0.000 .2306228 .5728186

_bs_3 | .66658 .0576837 11.56 0.000 .553522 .7796381

--

Mediation with multiple IVs

What if you had multiple independent variables? You just need to have one equation for each IV predicting the mediator variable. Here is the symbolic model.

sem (MV <- IV1)(MV <- IV2)(DV <- MV IV1 IV2)

For our example, we will use math and ses as our independent variables. We will keep the same mediator and dependent variable as before.

sem (read <- math)(read <- ses)(science <- read math ses) Endogenous variables

Observed: read science

Exogenous variables

Observed: math ses

Fitting target model:

Iteration 0: log likelihood = -2306.1661

Iteration 1: log likelihood = -2306.1661

Structural equation model Number of obs = 200

Estimation method = ml

Log likelihood = -2306.1661

--

| OIM

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

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

Structural |

read |

math | .68845 .059519 11.57 0.000 .5717949 .805105

ses | 1.726 .7698566 2.24 0.025 .2171093 3.234892

_cons | 12.43962 3.147394 3.95 0.000 6.270842 18.6084

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

science |

read | .3507374 .0663219 5.29 0.000 .2207487 .480726

math | .3905883 .0721193 5.42 0.000 .2492371 .5319395

ses | 1.033732 .731092 1.41 0.157 -.3991816 2.466647

_cons | 10.84415 3.065166 3.54 0.000 4.836532 16.85176

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

var(e.read)| 57.27968 5.727968 47.08476 69.68202

var(e.science)| 50.39009 5.039009 41.42142 61.30067

LR test of model vs. saturated: chi2(0) = 0.00, Prob > chi2 = .

We note that the indirect effects of both math and ses are significant.

Because we have multiple independent variables, the computation of the ratios and proportions is a bit more complex.

proportion of total math effect mediated = .2414651/.6320534 = .38203275
proportion of total ses effect mediated = .6053729/1.639105 = .36933137

ratio of math indirect to direct effect = .2414651/.3905883 = .61820874
ratio of ses indirect to direct effect = .6053729/1.033732 = .58561881

ratio of total math to direct effect = .6320534/.3905883 = 1.6182087
ratio of total ses to direct effect = 1.639105/1.033732 = 1.5856189

Mediation with multiple mediators

In this section we will consider the case in which there are multiple mediator variables.
This time there will be one equation for each mediator variable. The symbolic form of
the mode looks like this.

sem (MV1 <- IV)(MV2 <- IV)(DV <- MV1 MV2 IV)

For our example we will use read and write as the mediators.
We will go back to a single independent variable, math.

sem (read <- math)(write <- math)(science <- read write math) Endogenous variables

Observed: read write science

Exogenous variables

Observed: math

Fitting target model:

Iteration 0: log likelihood = -2779.4174

Iteration 1: log likelihood = -2779.4174

Structural equation model Number of obs = 200

Estimation method = ml

Log likelihood = -2779.4174

--

| OIM

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

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

Structural |

read |

math | .724807 .0579824 12.50 0.000 .6111636 .8384504

_cons | 14.07254 3.100201 4.54 0.000 7.996255 20.14882

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

write |

math | .6247082 .0562757 11.10 0.000 .5144099 .7350065

_cons | 19.88724 3.008947 6.61 0.000 13.98981 25.78467

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

science |

read | .3015317 .0679912 4.43 0.000 .1682715 .434792

write | .2065257 .0700532 2.95 0.003 .0692239 .3438274

math | .3190094 .0759047 4.20 0.000 .170239 .4677798

_cons | 8.407353 3.160709 2.66 0.008 2.212476 14.60223

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

var(e.read)| 58.71925 5.871925 48.26811 71.43329

var(e.write)| 55.31334 5.531334 45.46841 67.28993

var(e.science)| 48.77421 4.877421 40.09314 59.33492

LR test of model vs. saturated: chi2(1) = 21.43, Prob > chi2 = 0.0000

The indirect effect for math, .345706, is the combination of the indirect via
read plus the indirect via write. We can compute these indirect paths manually.

indirect via read = .724807*.3015317 = .21855229

indirect via write = .6247082*.2065257 = .1290183

total indirect = .724807*.3015317 + .6247082*.2065257 = .21855229 + .1290183 = .34757059

The last computation shows that the indirect effect given by estat teffects is the combined
indirect effect.

We can use the values we just computed to get the

ratios and proportions of interest.

proportion of total math effect mediated = .3475706/.66658 = .52142369
proportion of total math effect mediated via read = .21855229/.66658 = .32787106
proportion of total math effect mediated via write = .1290183/.66658 = .19355261

ratio of math indirect to direct effect = .3475706/.3190094 = 1.0895309
ratio of math indirect to direct effect via read = .21855229/.3190094 = .68509671
ratio of math indirect to direct effect via write = .1290183/.3190094 = .40443416

ratio of total math to direct effect = .66658/.3190094 = 2.0895309