What are the differences between mixed and sem linear growth models in Stata?

Authored by

stats writer

July 1, 2024

Mixed and sem linear growth models are two different statistical techniques used in Stata to analyze longitudinal or time-series data. Mixed models, also known as mixed-effects models or multilevel models, allow for the analysis of data with both fixed and random effects. This means that the model takes into account both individual-level differences and group-level differences in the data. On the other hand, sem linear growth models are a type of regression model that allows for the analysis of data with nonlinear relationships. This means that the model can capture nonlinear patterns in the data, such as quadratic or exponential growth. In Stata, mixed models are typically used for data with a hierarchical structure, while sem linear growth models are used for data with non-linear trends. Understanding the differences between these two models is important for choosing the appropriate approach for analyzing a specific dataset in Stata.

Linear growth models: mixed vs sem | Stata FAQ

Growth models are a very popular type of analysis. Many growth models can be run either with mixed or sem and yield the same results. This page will provide several examples of this.

We will begin by reading in the depression_clean dataset and changing it from wide into long form so that we can run mixed.

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

reshape long dep, i(sid) j(time)

(note: j = 0 1 2)
```

Data wide -> long

Number of obs. 46 -> 138

Number of variables 6 -> 5

j variable (3 values) -> time

xij variables:

dep0 dep1 dep2 -> dep

Unconditional growth model

We begin by running the unconditional growth model using mixed with
both random intercepts and random slope for time.

mixed dep time || sid:time, var cov(unstr)

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -414.27639
Iteration 1: log likelihood = -414.25833
Iteration 2: log likelihood = -414.25832

Computing standard errors:

Mixed-effects ML regression Number of obs = 138

Group variable: sid Number of groups = 46

Obs per group: min = 3
avg = 3.0
max = 3

Wald chi2(1) = 14.13
Log likelihood = -414.25832 Prob > chi2 = 0.0002

--

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

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

time | -1.6025 .4262612 -3.76 0.000 -2.437957 -.7670434

_cons | 14.18924 .8147121 17.42 0.000 12.59243 15.78605

--

--

Random-effects Parameters | Estimate Std. Err.

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

sid: Unstructured |
var(time) | 3.201386 2.047798 .9138158 11.21547
var(_cons) | 21.93819 6.613945 12.1501 39.61154
cov(time,_cons) | -1.153612 2.751286 -6.546034 4.23881

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

var(Residual) | 10.3135 2.15051 6.853596 15.52006

LR test vs. linear regression: chi2(3) = 54.85 Prob > chi2 = 0.0000

Next, we reshape the data back to wide and run the unconditional growth model using the
sem command. With this type of growth model we treat the intercept,
I and the slope, S as latent variables. We will follow
the convention that latent variable are in upper case while manifest variables are in
lower case.

reshape wide

(note: j = 0 1 2)

Data long -> wide

Number of obs. 138 -> 46
Number of variables 5 -> 6
j variable (3 values) time -> (dropped)
xij variables:

dep -> dep0 dep1 dep2

----

sem (dep0 <- I@1 S@0 _cons@0) ///
(dep1 <- I@1 S@1 _cons@0) ///
(dep2 <- I@1 S@2 _cons@0), ///
var(e.dep0@var e.dep1@var e.dep2@var) ///
means(I S)

Endogenous variables

Measurement: dep0 dep1 dep2

Exogenous variables

Latent: I S

Fitting target model:

Iteration 0: log likelihood = -418.88676
Iteration 1: log likelihood = -415.26423
Iteration 2: log likelihood = -414.28594
Iteration 3: log likelihood = -414.25861
Iteration 4: log likelihood = -414.25832
Iteration 5: log likelihood = -414.25832

Structural equation model Number of obs = 46

Estimation method = ml

Log likelihood = -414.25832

(1) I = 1

(2) I = 1

(3) S = 1

(4) I = 1

(5) S = 2

(6) _cons - _cons = 0

(7) _cons - _cons = 0

(8) _cons = 0

(9) _cons = 0

(10) _cons = 0

--

| OIM

| Coef. Std. Err. z Pgt;|z|

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

Measurement |

dep0 <- |

I | 1 (constrained)

_cons | 0 (constrained)

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

dep1 <- |

I | 1 (constrained)

S | 1 (constrained)

_cons | 0 (constrained)

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

dep2 <- |

I | 1 (constrained)

S | 2 (constrained)

_cons | 0 (constrained)

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

mean(I)| 14.18924 .814712 17.42 0.000 12.59243 15.78605

mean(S)| -1.6025 .4262611 -3.76 0.000 -2.437956 -.7670436

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

var(e.dep0)| 10.3135 2.150514 6.853595 15.52008

var(e.dep1)| 10.3135 2.150514 6.853595 15.52008

var(e.dep2)| 10.3135 2.150514 6.853595 15.52008

var(I)| 21.93818 6.613939 12.15009 39.61152

var(S)| 3.20138 2.047803 .913809 11.21551

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

cov(I,S)| -1.153606 2.751291 -0.42 0.675 -6.546037 4.238825

LR test of model vs. saturated: chi2(3) = 21.79, Prob > chi2 = 0.0001

Comparing the sem model with the mixed model shows that the parameter estimates are the same.

Time invariant covariate

Next, we will go back to the long form, run a mixed model adding a time invariant covariate, pre.

reshape long

(note: j = 0 1 2)

Data wide -> long

Number of obs. 46 -> 138
Number of variables 6 -> 5
j variable (3 values) -> time
xij variables:
dep0 dep1 dep2 -> dep

---

mixed dep time pre || sid:time, var cov(unstr)

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -411.12263

Iteration 1: log likelihood = -411.10613

Iteration 2: log likelihood = -411.10612

Computing standard errors:

Mixed-effects ML regression Number of obs = 138

Group variable: sid Number of groups = 46

Obs per group: min = 3

avg = 3.0

max = 3

Wald chi2(2) = 21.21

Log likelihood = -411.10612 Prob > chi2 = 0.0000

--

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

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

time | -1.6025 .4262611 -3.76 0.000 -2.437956 -.7670435

pre | .5051742 .1899545 2.66 0.008 .1328702 .8774781

_cons | 3.564548 4.073481 0.88 0.382 -4.419328 11.54842

--

--

Random-effects Parameters | Estimate Std. Err.

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

sid: Unstructured |

var(time) | 3.201384 2.047796 .9138156 11.21546

var(_cons) | 20.50672 6.374829 11.15031 37.71423

cov(time,_cons) | -2.289095 2.799971 -7.776937 3.198747

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

var(Residual) | 10.3135 2.15051 6.853597 15.52007

LR test vs. linear regression: chi2(3) = 45.83 Prob > chi2 = 0.0000

This last analysis is followed by its sem equivalent.

reshape wide

(note: j = 0 1 2)

Data long -> wide

Number of obs. 138 -> 46

Number of variables 5 -> 6

j variable (3 values) time -> (dropped)

xij variables:

dep -> dep0 dep1 dep2

```
sem (dep0 <- I@1 S@0 pre@p1 _cons@0) ///
(dep1 <- I@1 S@1 pre@p1 _cons@0) ///
(dep2 <- I@1 S@2 pre@p1 _cons@0), ///
var(e.dep0@var e.dep1@var e.dep2@var) ///
means(I S) covar(pre*I@0 pre*S@0)
```

Endogenous variables

Observed: dep0 dep1 dep2

Exogenous variables

Observed: pre

Latent: I S

Fitting target model:

Iteration 0: log likelihood = -563.45979 (not concave)

Iteration 1: log likelihood = -549.01197

Iteration 2: log likelihood = -538.31305

Iteration 3: log likelihood = -536.40749

Iteration 4: log likelihood = -536.3017

Iteration 5: log likelihood = -536.30149

Iteration 6: log likelihood = -536.30149

Structural equation model Number of obs = 46
Estimation method = ml
Log likelihood = -536.30149

(1) pre - pre = 0
(2) I = 1
(3) pre - pre = 0
(4) I = 1
(5) S = 1
(6) I = 1
(7) S = 2
(8) _cons - _cons = 0
(9) _cons - _cons = 0
(10) _cons = 0
(11) _cons = 0
(12) _cons = 0
(13) _cons = 0
(14) _cons = 0

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

Structural |

dep0 <- |

pre | .5051742 .1943431 2.60 0.009 .1242686 .8860797

I | 1 (constrained)

_cons | 0 (constrained)

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

dep1 <- |

pre | .5051742 .1943431 2.60 0.009 .1242686 .8860797

I | 1 (constrained)

S | 1 (constrained)

_cons | 0 (constrained)

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

dep2 <- |

pre | .5051742 .1943431 2.60 0.009 .1242686 .8860797

I | 1 (constrained)

S | 2 (constrained)

_cons | 0 (constrained)

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

Mean |

I | 3.564548 4.164044 0.86 0.392 -4.596828 11.72592

S | -1.6025 .4262611 -3.76 0.000 -2.437956 -.7670436

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

Variance |

e.dep0 | 10.3135 2.150514 6.853595 15.52008

e.dep1 | 10.3135 2.150514 6.853595 15.52008

e.dep2 | 10.3135 2.150514 6.853595 15.52008

I | 20.50671 6.374829 11.1503 37.71422

S | 3.20138 2.047803 .913809 11.21551

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

Covariance |

pre |

I | 0 (constrained)

S | 0 (constrained)

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

I |

S | -2.289091 2.79998 -0.82 0.414 -7.776951 3.198769

LR test of model vs. saturated: chi2(5) = 23.93, Prob > chi2 = 0.0002

Once again, the results are equivalent.

Time invariant covariate with cross-level interaction

This time we are going to add a cross-level interaction. Since, by now, you are accustomed to
the of reshape long, mixed, reshape wide
and sem, we will run everything in one long block of code and results.

Because we are predicting I and S with the time invariant covariate in the sem model, we can no longer request
mean(I S). These mean values will become parameters in the sem
output.

reshape long

(note: j = 0 1 2)

Data wide -> long

Number of obs. 46 -> 138
Number of variables 6 -> 5
j variable (3 values) -> time
xij variables:
dep0 dep1 dep2 -> dep
---

mixed dep c.time##c.pre || sid:time, var cov(unstr)

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -410.07935

Iteration 1: log likelihood = -410.05546

Iteration 2: log likelihood = -410.05544

Computing standard errors:

Mixed-effects ML regression Number of obs = 138

Group variable: sid Number of groups = 46

Obs per group: min = 3

avg = 3.0

max = 3

Wald chi2(3) = 24.02

Log likelihood = -410.05544 Prob > chi2 = 0.0000

--

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

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

time | -5.094745 2.417808 -2.11 0.035 -9.833561 -.3559284

pre | .3572517 .2150802 1.66 0.097 -.0642978 .7788012

|

c.time#c.pre | .1660464 .1132403 1.47 0.143 -.0559005 .3879933

|

_cons | 6.675614 4.592206 1.45 0.146 -2.324943 15.67617

--

--

Random-effects Parameters | Estimate Std. Err.

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

sid: Unstructured |
var(time) | 2.828174 1.981987 .7161158 11.16938
var(_cons) | 20.21054 6.267935 11.00507 37.11613
cov(time,_cons) | -1.95662 2.693749 -7.236271 3.32303

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

var(Residual) | 10.31349 2.150505 6.853593 15.52004

--

LR test vs. linear regression: chi2(3) = 46.84 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

reshape wide

(note: j = 0 1 2)

Data long -> wide

Number of obs. 138 -> 46

Number of variables 5 -> 6

j variable (3 values) time -> (dropped)

xij variables:

dep -> dep0 dep1 dep2

```
sem (dep0 <- I@1 S@0 _cons@0) ///
(dep1 <- I@1 S@1 _cons@0) ///
(dep2 <- I@1 S@2 _cons@0) ///
(I <- pre _cons) (S <- pre _cons), ///
var(e.dep0@var e.dep1@var e.dep2@var) ///
covar(e.I*e.S)
```

Endogenous variables

Measurement: dep0 dep1 dep2

Latent: I S

Exogenous variables

Observed: pre

Fitting target model:

Iteration 0: log likelihood = -836.11945 (not concave)

Iteration 1: log likelihood = -629.09569 (not concave)
Iteration 2: log likelihood = -572.06538 (not concave)
Iteration 3: log likelihood = -544.36594 (not concave)
Iteration 4: log likelihood = -540.10377
Iteration 5: log likelihood = -536.92737
Iteration 6: log likelihood = -535.30688
Iteration 7: log likelihood = -535.25089
Iteration 8: log likelihood = -535.25081
Iteration 9: log likelihood = -535.25081

Structural equation model Number of obs = 46
Estimation method = ml
Log likelihood = -535.25081

(1) I = 1
(2) I = 1
(3) S = 1
(4) I = 1
(5) S = 2
(6) _cons - _cons = 0
(7) _cons - _cons = 0
(8) _cons = 0
(9) _cons = 0
(10) _cons = 0

| OIM

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

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

Structural |

I <- |

pre | .3572517 .2150802 1.66 0.097 -.0642977 .7788011

_cons | 6.675614 4.592205 1.45 0.146 -2.324941 15.67617

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

S <- |

pre | .1660464 .1132402 1.47 0.143 -.0559003 .3879931

_cons | -5.094745 2.417806 -2.11 0.035 -9.833558 -.3559314

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

Measurement |

dep0 <- |

I | 1 (constrained)

_cons | 0 (constrained)

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

dep1 <- |

I | 1 (constrained)

S | 1 (constrained)

_cons | 0 (constrained)

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

dep2 <- |
I | 1 (constrained)
S | 2 (constrained)
_cons | 0 (constrained)
-------------+--
var(e.dep0)| 10.3135 2.150514 6.853595 15.52008
var(e.dep1)| 10.3135 2.150514 6.853595 15.52008
var(e.dep2)| 10.3135 2.150514 6.853595 15.52008
var(e.I)| 20.21051 6.267933 11.00505 37.11611
var(e.S)| 2.828156 1.981993 .716102 11.16945
-------------+--
cov(e.I,e.S)| -1.956604 2.693753 -0.73 0.468 -7.236263 3.323055

LR test of model vs. saturated: chi2(4) = 21.83, Prob > chi2 = 0.0002

Time-varying covariate

What if you have a time-varying covariate? We are going to switch datasets to lsay_long_clean
to show an example with a time varying covariate, att.

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

clear

mixed math c.yr c.att || id:yr, var cov(unstr)

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -36146.122
Iteration 1: log likelihood = -36144.71
Iteration 2: log likelihood = -36144.708

Computing standard errors:

Mixed-effects ML regression Number of obs = 10785
Group variable: id Number of groups = 3595

Obs per group: min = 3
avg = 3.0
max = 3

Wald chi2(2) = 2340.50
Log likelihood = -36144.708 Prob > chi2 = 0.0000

--

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

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

yr | 2.64315 .0546525 48.36 0.000 2.536033 2.750267

att | .1700024 .0253111 6.72 0.000 .1203936 .2196112

_cons | 54.67699 .3330636 164.16 0.000 54.0242 55.32978

Random-effects Parameters | Estimate Std. Err.

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

id: Unstructured |

var(yr) | 3.348592 .3030205 2.804371 3.998427

var(_cons) | 110.5491 2.912331 104.9859 116.4071

cov(yr,_cons) | -.0107825 .6369843 -1.259249 1.237684

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

var(Residual) | 14.50231 .3427178 13.84592 15.18983

LR test vs. linear regression: chi2(3) = 10678.18 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

Back to the old drill of reshaping wide and running a sem model.

This model proved to be a bit fussier and required that we provide starting values for the coefficients. To obtain proper starting values we ran a simpler model and saved the results into a matrix. We then used these results as starting values for the full model.

reshape wide math att, i(id) j(yr)

(note: j = 0 1 2)

Data long -> wide

Number of obs. 10785 -> 3595
Number of variables 7 -> 10
j variable (3 values) yr -> (dropped)
xij variables:
math -> math0 math1 math2
att -> att0 att1 att2

---

sem (math0 <- I@1 S@0 _cons@0) ///
(math1 <- I@1 S@1 _cons@0) ///
(math2 <- I@1 S@2 _cons@0), ///
var(e.math0@var e.math1@var e.math2@var) ///

means(I S)

mat b = e(b)

sem (math0 <- I@1 S@0 att0@b1 _cons@0) ///
(math1 <- I@1 S@1 att1@b1 _cons@0) ///
(math2 <- I@1 S@2 att2@b1 _cons@0), ///
var(e.math0@var e.math1@var e.math2@var) ///
means(I S) covar(att0*I@0 att1*I@0 att2*I@0) ///
covar(att0*S@0 att1*S@0 att2*S@0) ///
from(b)

Endogenous variables

Observed: math0 math1 math2

Exogenous variables

Observed: att0 att1 att2
Latent: I S

Fitting target model:

Iteration 0: log likelihood = -61901.22
Iteration 1: log likelihood = -60959.753
Iteration 2: log likelihood = -60758.068

Iteration 3: log likelihood = -60746.189

Iteration 4: log likelihood = -60746.116

Iteration 5: log likelihood = -60746.116

Structural equation model Number of obs = 3,595

Estimation method = ml

Log likelihood = -60746.116

(1) att0 - att2 = 0

(2) I = 1

(3) att1 - att2 = 0

(4) I = 1

(5) S = 1

(6) I = 1

(7) S = 2

(8) _cons - _cons = 0

(9) _cons - _cons = 0

(10) _cons = 0

(11) _cons = 0

(12) _cons = 0

| OIM

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

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

Structural |

math0 <- |

att0 | .1700025 .025449 6.68 0.000 .1201234 .2198816

I | 1 (constrained)

_cons | 0 (constrained)

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

math1 <- |

att1 | .1700025 .025449 6.68 0.000 .1201234 .2198816

I | 1 (constrained)

S | 1 (constrained)

_cons | 0 (constrained)

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

math2 <- |

att2 | .1700025 .025449 6.68 0.000 .1201234 .2198816

I | 1 (constrained)

S | 2 (constrained)

_cons | 0 (constrained)

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

mean(I)| 54.67699 .3343215 163.55 0.000 54.02173 55.33225

mean(S)| 2.64315 .0546563 48.36 0.000 2.536026 2.750275

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

var(e.math0)| 14.50234 .3427203 13.84594 15.18986

var(e.math1)| 14.50234 .3427203 13.84594 15.18986

var(e.math2)| 14.50234 .3427203 13.84594 15.18986

var(I)| 110.5491 2.91233 104.9859 116.4071

var(S)| 3.348555 .3030222 2.804331 3.998394

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

cov(I,S)| -.0107522 .6369845 -0.02 0.987 -1.259219 1.237714

--

LR test of model vs. saturated: chi2(11) = 201.05, Prob > chi2 = 0.0000

We hope this helps get you started with linear growth models.