Multilevel Models with Two Levels using Stata 15 – Guided Practical (2)

1) Open dataset *tutorial.dta*

We will be modelling student attainment (*normexam*) as a linear function of gender (*girl*), ability (*standlrt*), and several other variables that are specified below.

The variable *normexam* records the student's scores in examinations at age 16, normalised to have a mean of 0 and a standard deviation of 1. The variable *standlrt* is the student's result on a reading test at age 11 also standardised to have a mean of 0 and a standard deviation of 1.

2) First run a VC model and a random intercept with only one explanatory at a time. We will use the regression coefficients and residual estimates to compare models.

mixed normexam ||school:student, ml variance

3) Also run a single-level model. We will compare the results with the 2-level models.

mixed normexam || , ml variance

4) Run a two-level model with the variable 'girl'.

mixed normexam girl || school:student, ml variance

and explore how it differs from a single-level model

mixed normexam girl || , ml variance

5) Run the same analysis with only the explanatory variable prior ability ('standlrt').

mixed normexam standlrt || school:student, ml variance

6) We will now look at the simultaneous effect of gender and prior attainment on normexam:
mixed normexam girl standlrt || school:student, ml variance

```
Mixed-effects ML regression                     Number of obs     =      4,059
Group variable: school                          Number of groups  =         65

                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198

                                                Wald chi2(2)      =    2084.36
Log likelihood = -4665.0033                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
    normexam |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        girl |   .1713752   .0327609     5.23   0.000      .107165    .2355853
    standlrt |   .5595383   .0124479    44.95   0.000     .5351408    .5839357
       _cons |  -.0949117   .0434129    -2.19   0.029    -.1799993   -.009824
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Independent          |
                var(student) |   1.94e-21   1.14e-20      1.85e-26    2.02e-16
                  var(_cons) |   .0880751   .0177868       .059286     .130844
-----------------------------+------------------------------------------------
               var(Residual) |   .5622563   .0125849      .5381237    .5874712
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 387.00                Prob > chi2 = 0.0000
```

7) The next step is to investigate whether the association between prior attainment and attainment at age 16 is different for boys and girls.

To run two separate regression models (one for boys and one for girls), type:

mixed normexam girl standlrt if girl==1 || school:student, ml variance
mixed normexam girl standlrt if girl==0 || school:student, ml variance
The resulting regression coefficients of *normexam* on *standlrt* are:

girls: 0.556 (0.016)

boys: 0.565 (0.019)

But that is far too cumbersome, and we can do it more neatly by inserting an **interactive effect** in a single model.

To set up an interactive effect of two explanatory variables, type:

*gen girlstandlrt = girl*standlrt*

and then add this term to the model, by typing

mixed normexam girl standlrt girlstandlrt|| school:student, ml variance

```
Mixed-effects ML regression                     Number of obs     =      4,059
Group variable: school                          Number of groups  =         65

                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198

                                                Wald chi2(3)      =    2084.44
Log likelihood = -4664.9776                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
    normexam |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        girl |   .1712928   .0327628     5.23   0.000      .107079    .2355066
    standlrt |   .5626053   .0183701    30.63   0.000     .5266006    .5986101
girlstandlrt |  -.0055797   .0245754    -0.23   0.820    -.0537467    .0425873
       _cons |  -.0946898   .0434248    -2.18   0.029    -.1798008   -.0095788
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Independent          |
                var(student) |   1.65e-21   1.05e-20      6.11e-27    4.46e-16
                  var(_cons) |   .0880802   .0177879      .0592894    .1308516
-----------------------------+------------------------------------------------
               var(Residual) |   .5622486   .0125847      .5381163    .5874632
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 387.03                Prob > chi2 = 0.0000
```

The results tell us that the slope of *normexam* on *standlrt* is:

0.563 when girl=0;

0.563 – 0.006, which is 0.557, when girl=1.

This is consistent with the two separate regressions: the slope is shallower for girls than for boys, but the difference *0.006* is negligible because its standard error is *0.025* (and thus the t-value is only *0.23*).

8) We can also add explanatory variables at the group level. Consider the variable *schgend*, which records whether the school is mixed-gender, boys-only or girls-only. Because the data set specifies this as a categorical variable, you should include the variable with the prefix **i.schgend**. The reference category is by default the first, which here is **mixedsch**. If you wish to specify a different reference category, specify the number of the category in the prefix, e.g. **ib2.schgend**.

In technical detail, what is being done here is to define two *dummy variables*:

boysch = 1 for pupils in boys-only schools, and 0 for all other pupils;

girlsch = 1 for pupils in girls-only schools, and 0 for all other pupils.

The pupils who have value 0 on both of these – that is all pupils who are in mixed schools – are therefore the reference group. The estimated regression coefficients of *boysch* and *girlsch* will then be estimates of the difference from that reference group of (respectively) boys in boys-only schools and girls in girls-only schools.

To run the model, type
mixed normexam girl standlrt i.schgend|| school:student, ml variance

```
Mixed-effects ML regression                     Number of obs     =      4,059
Group variable: school                          Number of groups  =         65

                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198

                                                Wald chi2(4)      =    2093.27
Log likelihood = -4662.7132                     Prob > chi2       =     0.0000
```

normexam	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
girl	.1672282	.0340818	4.91	0.000	.100429	.2340273
standlrt	.5599641	.0124436	45.00	0.000	.5355752	.5843531
schgend						
boysch	.1776197	.1107534	1.60	0.109	-.0394529	.3946923
girlsch	.1589596	.0872548	1.82	0.068	-.0120567	.3299759
_cons	-.1681504	.0539994	-3.11	0.002	-.2739873	-.0623134

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
school: Independent				
var(student)	2.19e-19	1.09e-18	1.32e-23	3.65e-15
var(_cons)	.0811077	.0165468	.0543761	.1209807
var(Residual)	.5622731	.0125854	.5381393	.5874891

```
LR test vs. linear model: chi2(2) = 346.77                Prob > chi2 = 0.0000
```

Look at the results of the new variable ‘schgend’.

boys’ schools: 0.178 (0.111); t-value = 1.6

girls’ schools: 0.159 (0.088); t-value = 1.8

There is no compelling evidence that ability scores differ by school gender.

4) In a formal sense, adding interactive effects between variables measured at two different levels is exactly the same as adding them when they are at the same level (as we did in section **Error! Reference source not found.** for *girl* and *standlrt*). For example, when we add the interactive effect of *standlrt* and *schgend*, we are testing whether the effect of ability on attainment is different in the different kinds of school.

To run a two-level model including an interaction term between a categorical and a continuous variable, we type:

mixed normexam c.standlrt girl i.schgend c.standlrt#i.schgend||school:student, ml variance

The results for the slope connecting ability and attainment are shown below.

```
Mixed-effects ML regression                     Number of obs     =      4,059
Group variable: school                          Number of groups  =         65

                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198

                                                Wald chi2(6)      =    2094.01
Log likelihood = -4662.4647                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------------
           normexam |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
           standlrt |   .5671806   .0171576    33.06   0.000     .5335522    .6008089
               girl |   .1660151   .0341376     4.86   0.000     .0991065    .2329236
                    |
            schgend |
            boysch  |   .1769647   .1108121     1.60   0.110    -.0402231    .3941525
           girlsch  |   .1593937   .0872701     1.83   0.068    -.0116527      .33044
                    |
schgend#c.standlrt  |
            boysch  |   -.005859   .0364759    -0.16   0.872    -.0773504    .0656324
           girlsch  |  -.0195373   .0277719    -0.70   0.482    -.0739693    .0348947
                    |
              _cons |  -.1673906   .0540208    -3.10   0.002    -.2732693   -.0615118
------------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Independent          |
               var(student)  |   3.16e-19   1.92e-18      2.19e-24    4.57e-14
                 var(_cons)  |   .0811363   .0165528      .0543951    .1210237
-----------------------------+------------------------------------------------
              var(Residual)  |   .5622004   .0125838      .5380698    .5874132
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 347.12                Prob > chi2 = 0.0000
```

This tells us that the baseline slope (i.e. in mixed schools) is *0.567* (s.e. *0.017*).

The slope in boys' schools is *0.567 – 0.006 = 0.561*, and the slope in girls' schools is *0.567 – 0.020 = 0.547*. Neither of these differences in slope (*0.006* and *0.020*) is even as large as its standard error, and so there is no evidence of different slopes. More formally, we test this by comparing Deviance values: *347.12* in this model, compared to *346.77* in the model without the interactive effect, a difference of only 0.35 on 2 degrees of freedom (since we have estimated 2 extra regression parameters – the two different slopes). So, this Deviance test confirms that there is no evidence of a difference in slopes.

However, another cross-level interactive effect shows clear evidence of difference – this one between the effect of the student-level ability score *standlrt* and the effect of the average ability score in the school, *avslrt*. What this interactive effect is testing is whether the association between ability and attainment varies according to the average ability of the students in the school.

mixed normexam c.standlrt c.avslrt girl c.standlrt#c.avslrt||school:student, ml variance

***Note that in addition to the interaction term, we need to add separately the two variables that form our interaction term ***

```
Mixed-effects ML regression                     Number of obs     =      4,059
Group variable: school                          Number of groups  =         65

                                                Obs per group:
                                                              min =          2
                                                              avg =       62.4
                                                              max =        198

                                                Wald chi2(4)      =    2134.23
Log likelihood = -4650.5203                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------------
          normexam |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
          standlrt |   .5600309   .0125259    44.71   0.000     .5354805    .5845812
            avslrt |   .3542316   .1076468     3.29   0.001     .1432477    .5652154
              girl |   .1703175   .0324361     5.25   0.000      .106744     .233891
                   |
c.standlrt#c.avslrt |   .1738331   .0389759     4.46   0.000     .0974417    .2502245
                   |
             _cons |  -.1044247   .0406962    -2.57   0.010    -.1841877   -.0246617
------------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school: Independent          |
                var(student) |   3.98e-23   1.99e-22      2.27e-27    6.99e-19
                  var(_cons) |   .0718964   .0150764      .0476665     .108443
-----------------------------+------------------------------------------------
               var(Residual) |   .5598074   .0125331      .5357742    .5849187
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 319.00                Prob > chi2 = 0.0000
```

The interactive term – a value of *0.174* with standard error of *0.039* (and thus t-value of 4.5) – is strongly statistically significant. It means that the higher the average ability in the school, the steeper the slope connecting attainment and ability.

To get a sense of what this means, note that the school-average variable *avslrt* has a standard deviation in the sample of *0.315 (sum avslrt)*. So, in a school that has above-average ability (specifically, one standard deviation above the mean), the expected slope of attainment on individual ability is *0.56 + 0.174x0.315 = 0.615*. Similarly, for a school with below-average ability (one standard deviation below the mean) the slope would be *0.56 – 0.174x0.315 = 0.505*.

Recall that *standlrt* itself has standard deviation of about 1. Thus, the gap in expected attainment between students 1 standard deviation apart is twenty percent greater in a school with, in this sense, high average ability than it is in schools with below-average ability (0.615 compared to 0.505).