Multilevel Models with Two Levels using Stata 15 – Guided Practical

During this session we will:

1) Show you how to run a multilevel model in Stata.
2) Show you how to add several explanatory variables to the model.
3) Show you how to include interactive effects between explanatory variables.
4) Show you how to include and interpret explanatory variables that are measured at the group level.
5) Show you how to include and interpret interactive effects between explanatory variables that are measured at different levels.

We work with the hedonism data from the ESS. The data have a two-level hierarchical structure with individual respondents at level 1 and countries at level 2. We will treat country as a random classification. The target of inference could be a wider population of countries from which those in the study can be considered a random sample. However, it is not clear which countries such a population would contain. In this case, it is more natural to think of the sample data as if they were a set of realisations from some underlying process that could extend through time and possibly space[1]. This process has driven the observations, but the statistics we compute from the observed data refer to a particular point in time and are subject to random fluctuations. We are interested in the underlying process that has generated the data we observe, and use the 'sample' data to make inferences about this process.

We will be using four variables:

- age - Respondent's age in years
- female 0 if respondent male, 1 if respondent female
- eduyrs - Number of years of education of respondent
- income - Respondent's monthly household income in bands (less than €150, €150-300, €300-500, €500-1000, €1000-1500, €1500-2000, €2000-2500, €2500-3000, €3000-5000, €5000-7500, €7500-10000, more than €10000).

We will analyse data from all 20 countries in the study. The following countries were included in the study: Austria, Belgium, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom. The combined sample size for these countries is 36,537.

[1] In survey sampling this abstract notion of a target population is called a *superpopulation*. A superpopulation is infinite, while a population consisting of a fixed number of countries (e.g. all European countries) is finite.

1) We will start by opening the dataset **hedon_Random Intercept.dta** in Stata 15 and study the variables, using commands like *sum, des, codebook* and *tab*.

Higher scores on the hedonism variable indicate more hedonistic beliefs.

describe

codebook

sum hed country

histogram hed, frequency

tab country

```
. describe age female eduyrs income

              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------
age             double  %9.0g
female          double  %9.0g
eduyrs          double  %9.0g
income          double  %9.0g
```

2) We will continue by fitting simplest multilevel model which allows for country effects on hedonism, but with no explanatory variables – **a *variance components model***. This model may be written

$$y_{ij} = \beta_0 + u_j + e_{ij} \qquad (1)$$

where y_{ij} is the hedonism score of respondent i in country j, β_0 is the overall mean across countries, u_j is the effect of country j on hedonism, and e_{ij} is a person-level residual. The country effects u_j, which we will also refer to as country (or level 2) residuals, are assumed to follow a normal distribution with mean zero and variance σ_u^2.

To run this model in Stata 15, the command is

mixed y x || lev2: , ml

The default is maximum likelihood (ML) estimation. *Variance* show random-effects and residual-error parameter estimates as variances and covariances.

For this example type:

mixed hed || country:, ml variance

```
. mixed hed|| country:, ml variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -49651.675
Iteration 1:   log likelihood = -49651.675  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =     36,527
Group variable: country                         Number of groups  =         20

                                                Obs per group:
                                                              min =      1,213
                                                              avg =    1,826.3
                                                              max =      2,785

                                                Wald chi2(0)      =          .
Log likelihood = -49651.675                     Prob > chi2       =          .
```

hed	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
_cons	-.2031517	.0671841	-3.02	0.002	-.3348301	-.0714732

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
country: Identity				
var(_cons)	.0897673	.0285524	.0481255	.1674407
var(Residual)	.8850616	.0065509	.8723147	.8979947

```
LR test vs. linear model: chibar2(01) = 3286.25        Prob >= chibar2 = 0.0000
```

3) We will now calculate the **Variance Partition coefficient**:

First, calculate the total variance = $\sigma_u^2 + \sigma_e^2 =$,

Second, the VPC = $\frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} =$

Amount (calculate the percentage) of variance in hedonism explained by differences between countries.

4) Test the significance of country effects. We shall do that by carrying out a likelihood ratio test comparing the variance components model with the corresponding single-level model (also without explanatory variables). To fit the single-level model, we need to remove the random country effects:

mixed hed || , ml variance

```
. mixed hed|| , ml variance

Mixed-effects ML regression                     Number of obs     =     36,527

                                                Wald chi2(0)      =          .
Log likelihood =   -51294.8                     Prob > chi2       =          .
```

hed	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	-.2035884	.0051563	-39.48	0.000	-.2136944	-.1934823

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
var(Residual)	.9711439	.0071861	.9571611	.9853309

The likelihood ratio test statistic is calculated as the difference in the log likelihood values for the two models:

> *LR = 2(51294.8 – 49651)= 3288 on 1 d.f. (because there is only parameter difference between the models,* σ_{u0}^2*).*

Bearing in mind that the 5% point of a chi-squared distribution on 1 d.f. is 3.84, there is overwhelming evidence of country effects on hedonism. We will therefore revert to the multilevel model with country effects.

Another way of doing that in Stata 15:

qui mixed hed || country:, ml variance

estimates store vc

qui mixed hed || , ml variance

estimates store sl

lrtest vc sl

5) We will now calculate the **random intercepts model**. We will look at whether these differences between countries do in fact remain after controlling for a range of individual characteristics. We will initially just consider individual age, but in section 2 we will add in individual income, education and gender. In order to address our question as to whether differences in hedonism between countries remain after controlling for individual age (or for individual age, income, education and gender), we will need to fit a random intercepts model, by adding an explanatory variable (or variables) into our variance components model.

Firstly, **run a variance component model** and store the results of the model, so that we will be able to compare the results easily with the results of the random intercepts model we are about to fit.

mixed hed|| country: , ml variance

estimates store vc

6) We will now fit the random intercepts model. We get a random intercepts model by simply adding an explanatory variable in the variance component model. We are initially trying to answer the question `Do differences in hedonism between countries remain after controlling for individual age?' and so we will add age as our explanatory variable. We will centre age so that the intercept will be easier to interpret

Centre age around the mean:
egen meanage = mean(age)

gen centage = age - meanage

Run Random Intercepts model:

mixed hed centage|| country: , ml variance

estimates store ri

```
. mixed hed age2|| country: , ml variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -47294.541
Iteration 1:   log likelihood = -47294.541  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =     36,364
Group variable: country                         Number of groups  =         20

                                                Obs per group:
                                                              min =      1,202
                                                              avg =    1,818.2
                                                              max =      2,770

                                                Wald chi2(1)      =    4549.62
Log likelihood = -47294.541                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
         hed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age2 |  -.0174069   .0002581   -67.45   0.000    -.0179127   -.0169011
       _cons |  -.1989077   .0670277    -2.97   0.003    -.3302796   -.0675358
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
country: Identity            |
                  var(_cons) |   .0894017   .0284152      .0479515    .1666822
-----------------------------+------------------------------------------------
               var(Residual) |   .7869226   .0058376      .7755639    .7984475
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 3730.57       Prob >= chibar2 = 0.0000
```

To compare the results of the two models:

estimates table vc ri

The command *estat vce*, gives you the variances and covariances of the models.

```
. estimates table vc ri
```

Variable	vc	ri
hed		
age2		-.0174069
_cons	-.20315166	-.1989077
lns1_1_1		
_cons	-1.2052672	-1.2073077
lnsig_e		
_cons	-.06104903	-.11981272

The coefficient for age is -0.017 and can be interpreted as follows. Older people are less hedonistic; for every extra year, **hed** drops by 0.017.

Compared to the variance components model, the intercept has hardly changed, and nor has the level 2 variance. Since the level 2 variance is still large it appears that the answer to our question `Do differences in hedonism between countries remain after controlling for individual age?' is `Yes'. However, we cannot be sure of this until we have tested the significance of the level 2 variance σ^2_{u0}.

The level 1 variance has decreased, from 0.885 to 0.787. Thus, some of the variation we observed between individuals within each country in the variance components model turns out to have been due to variation in age between the individuals, and the relationship between age and hedonism.

A substantial amount of level 1 variation remains nevertheless; and the level 1 variation is still much greater than the level 2 variation. There must be other factors which determine how hedonistic an individual is.

7) We will now test the significance of σ^2_{u0}, in order to answer our question `Do differences in hedonism between countries remain after controlling for individual age?'. We use a likelihood ratio test. We will be comparing the log likelihood of our random intercepts model with the log likelihood of a single level model with the same explanatory variables, so we'll need to fit that single level model to obtain its log likelihood.

Run and store a single level model:

mixed hed centage|| , ml variance
estimates store SL1

```
. mixed hed age2|| , ml variance

Mixed-effects ML regression                     Number of obs     =     36,364

                                                Wald chi2(1)      =    4032.11
Log likelihood = -49159.825                     Prob > chi2       =     0.0000
```

hed	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age2	-.0171544	.0002702	-63.50	0.000	-.0176839	-.0166249
_cons	-.2010064	.0049041	-40.99	0.000	-.2106182	-.1913945

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
var(Residual)	.8744906	.0064854	.8618714	.8872945

The model comparison table should now look like this:

estimates vc ri SL1

```
. estimates table vc ri SL1
```

Variable	vc	ri	SL1
hed			
age2		-.0174069	-.01715443
_cons	-.20315166	-.1989077	-.20100636
lns1_1_1			
_cons	-1.2052672	-1.2073077	
lnsig_e			
_cons	-.06104903	-.11981272	-.06705687

Our test statistic is 2(log(likelihood(SL1)) - log(likelihood(ri))), i.e. 98319.650 – 94589.109 = 3730.541. This is a large value for the test statistic, so we can reject the null hypothesis that $\sigma_{u0}^2 = 0$ at all conventionally used levels of significance. We can conclude that differences between countries do indeed remain after controlling for age.

Looking at the random intercept model, is the coefficient for age significant? Its value is -0.017 and its standard error 0.000. To test whether it is significant at the 5% level we must divide the estimate by the standard error and compare the absolute value of the result with 2 (or, more precisely, with 1.96). The value of the standard error is never equal to 0, it is some very small value that might be 0.000 when rounded to the 3 decimals places. In other words, it is less than 0.0005. So here we divide our coefficient by 0.0005. -0.017 divided by 0.0005 is -34. The absolute value 34 > 1.96 which implies that the coefficient is significant at the 5% level.

8) We will now expand our question to ask `Do differences in hedonism between countries remain after controlling for individual age, income, education and gender?'. To answer this we will add these explanatory variables. We'll treat income, technically an ordinal variable, as continuous since it has a large number of categories (making interpretation cumbersome if entered as dummies) and its relationship with hedonism is fairly linear (as exploration of the data would show), so that this appears to be a valid way of handling it. We will again centre the continuous variables round their means.

First, we need to centre our explanatory variables:

egen meanincome=mean(income)
gen income2 = income - meanincome
sum eduyrs
egen meanedu=mean(eduyrs)
gen edu2=eduyrs - meanedu

Also, use the variable *centage* (age centred).

And then we run our random intercept model:

mixed hed centage income2 edu2 female || country: , ml variance

```
. mixed hed age2 income2 edu2 female || country: , ml variance

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -38321.519
Iteration 1:   log likelihood = -38321.519  (backed up)

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =     29,419
Group variable: country                         Number of groups  =         20

                                                Obs per group:
                                                              min =        875
                                                              avg =    1,471.0
                                                              max =      2,223

                                                Wald chi2(4)      =    3555.05
Log likelihood = -38321.519                     Prob > chi2       =     0.0000
```

hed	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age2	-.0174937	.0003139	-55.73	0.000	-.018109	-.0168785
income2	.0088879	.002783	3.19	0.001	.0034333	.0143426
edu2	-.0066012	.0015214	-4.34	0.000	-.0095831	-.0036194
female	-.1112955	.0104358	-10.66	0.000	-.1317493	-.0908417
_cons	-.1463267	.066664	-2.19	0.028	-.2769857	-.0156677

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
country: Identity				
var(_cons)	.0877163	.0279385	.0469856	.1637557
var(Residual)	.789698	.0065134	.7770346	.8025679

```
LR test vs. linear model: chibar2(01) = 2749.78          Prob >= chibar2 = 0.0000
```

Looking at the coefficient of age and carrying out the same procedure as above we can see that it remains negative (with the same value) and significant at the 5% level after adding in eduyrs, income and female.

Are income, eduyrs and female significant at the 5% level? Yes, the coefficients of all these variables are more than twice their standard errors so they are all significant at the 5% level.

What are the relationships between hedonism and income, hedonism and education, and hedonism and gender? People with more education are less hedonistic (hed drop by 0.007 for every additional years of education), people with higher income are more hedonistic (hed rises by 0.009 for every income band risen), and women are less hedonistic than men (hed is 0.111 lower for women than for men).

Has there been any change in the estimates for the variances in this model compared to the first random intercept model? The level 1 and level 2 variances have both remained roughly the same compared to the first random intercept model (model ri). The small differences between the two models may be partly due to the fact that the models are fitted to slightly different samples. In fact, if you look at the outputs you can see that 29,419 observations were used in the second model, compared to 36,364 observations that were used for the first model. Individuals with missing values for any of the explanatory variables are not included in that model. Thus, there were around 7000 observations that were used in model RI 1 but not in RI 2. Note that when performing a likelihood ratio test it is important to ensure that the models you are comparing are fitted to the same sample, so for example we couldn't use a likelihood ratio test to compare the first and the second random intercept models. It is important to check that the sample is the same for both models when using a likelihood ratio test.

Do differences in hedonism between countries remain after controlling for individual age, income, education and gender? To answer this question we need to test whether σ^2_{uo} is significant. From the output above we see that the test statistic is 2749.78, with a p-value of 0.0000. We reject the null hypothesis and conclude that σ^2_{uo} is significant at 5% level. So, differences in hedonism between countries do remain after controlling for individual age, income, education and gender.

9) We will now look whether the level 2 variance has become larger or smaller after adding these variables, and whether the level 2 variance accounts for a larger or smaller proportion of the total variance after adding these variables.

To see whether the level 2 variance accounts for a larger or smaller proportion of the total variance in models RI 1 and RI 2 compared to model VC we need to calculate the VPC. We'll return to the above outputs to get the parameter estimates we need to do this.

For the VC model the, VPC is $\rho = \frac{0.090}{0.090+0.885} = 9.2\%$

For the RI 1 model the, VPC is $\rho = \frac{0.090}{0.090+0.787} = 10.3\%$

For the RI 2 model the, VPC is $\rho = \frac{0.088}{0.088+0.79} = 10\%$

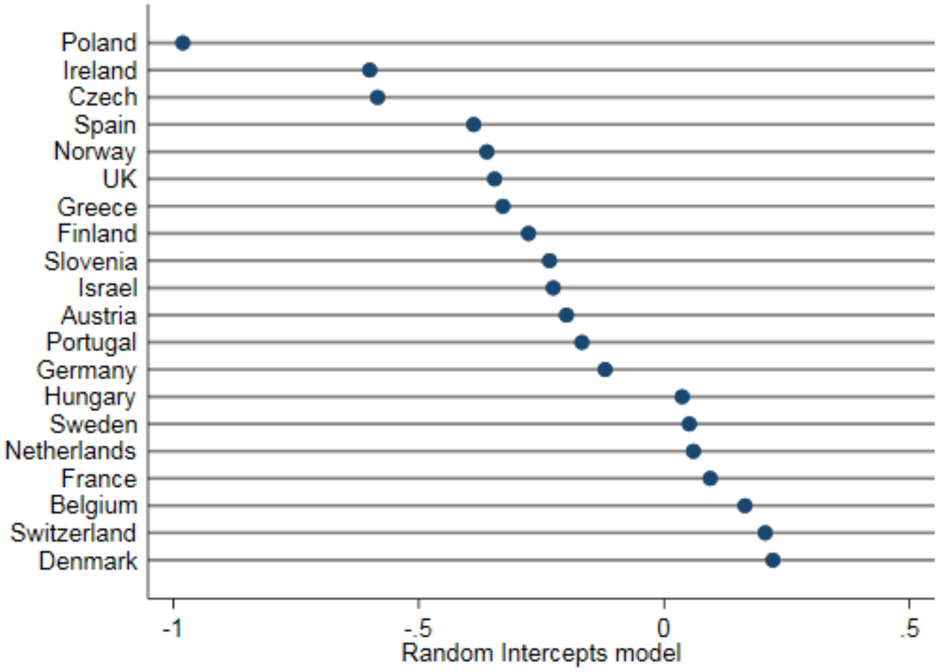
We can conclude that introducing individual age, income, education and gender did not account for much of the differences between countries in hedonism, nor did it unmask further differences between countries in hedonism.

In each of the random intercept models there is a greater proportion of variation at the country level than in the variance components model. This is as we would expect since we noted that when we added in our explanatory variables the country level variation remained almost unchanged while the individual level variation dropped compared to the variance components model. This means that after accounting for individual age (or individual age, income, education and gender) the differences in hedonism between countries become more important. (Remember that we are talking about the importance of differences between countries for the unexplained variation, and that we have just reduced that unexplained variation by accounting for some differences in hedonism due to individual characteristics). However, we observe very small differences in the VPC from the three models, which would lead us to think that the importance of countries in explaining variation in hedonism remains about the same after accounting for individual age, income, education and gender.

10) We will now calculate and plot the residuals for the models

To calculate and plot the level 2 residuals, type:
predict r1, reffects
des r1
gen b0=_b[_cons] + r1
bysort country: gen tolist = _n==1
list country b0 if tolist
graph dot (mean) b0, over(country, gap(country) sort(b0)) cw linetype(line) ytitle(Random Intercepts model)

	country	b0
1.	Austria	-.1989642
2204.	Belgium	.1645687
4023.	Switzerla	.2058058
6028.	Czech	-.5840442
7241.	Germany	-.1204819
10026.	Denmark	.221467
11481.	Spain	-.3880341
13119.	Finland	-.2764339
14875.	France	.0938309
16187.	UK	-.3455155
17935.	Greece	-.3286789
20388.	Hungary	.0366108
21952.	Ireland	-.5999266
23790.	Israel	-.2261041
25957.	Netherlan	.059483
28258.	Norway	-.3612163
30064.	Poland	-.980716
32044.	Portugal	-.1675834
33461.	Sweden	.051253
35138.	Slovenia	-.233474



Studying the table and graph above, we can find out which country is the most hedonistic after controlling for individual age.

Also, we can compare the level-2 residuals by country of the two models to see whether the residuals have roughly stayed the same or whether the residuals for the individual countries have changed while only the overall pattern remains the same.

Variance Components Model

	country	b0
1.	Austria	-.1886473
2204.	Belgium	.1871823
4023.	Switzerla	.182179
6028.	Czech	-.6789357
7241.	Germany	-.1279914
10026.	Denmark	.2192593
11481.	Spain	-.4296367
13119.	Finland	-.2646315
14875.	France	.1056406
16187.	UK	-.3831785
17935.	Greece	-.3754116
20388.	Hungary	.042697
21952.	Ireland	-.5750204
23790.	Israel	-.1526639
25957.	Netherlan	.0285833
28258.	Norway	-.3645766
30064.	Poland	-.9262211
32044.	Portugal	-.198098
33461.	Sweden	.0363296
35138.	Slovenia	-.1998915

Random intercepts Model

	country	b0
1.	Austria	-.1989642
2204.	Belgium	.1645687
4023.	Switzerla	.2058058
6028.	Czech	-.5840442
7241.	Germany	-.1204819
10026.	Denmark	.221467
11481.	Spain	-.3880341
13119.	Finland	-.2764339
14875.	France	.0938309
16187.	UK	-.3455155
17935.	Greece	-.3286789
20388.	Hungary	.0366108
21952.	Ireland	-.5999266
23790.	Israel	-.2261041
25957.	Netherlan	.059483
28258.	Norway	-.3612163
30064.	Poland	-.980716
32044.	Portugal	-.1675834
33461.	Sweden	.051253
35138.	Slovenia	-.233474

We can see from the table above that there is only a small amount of change in the ranking of the countries in the random intercepts model compared to the variance components model. Countries that have a comparatively high level of hedonism still have a comparatively high level of hedonism after controlling for the age of the respondents. This is not a surprising finding given that the country level variance did not change much after adding age.

11) We will now the model assumptions by checking the normality of level 1 and level 2 residuals.

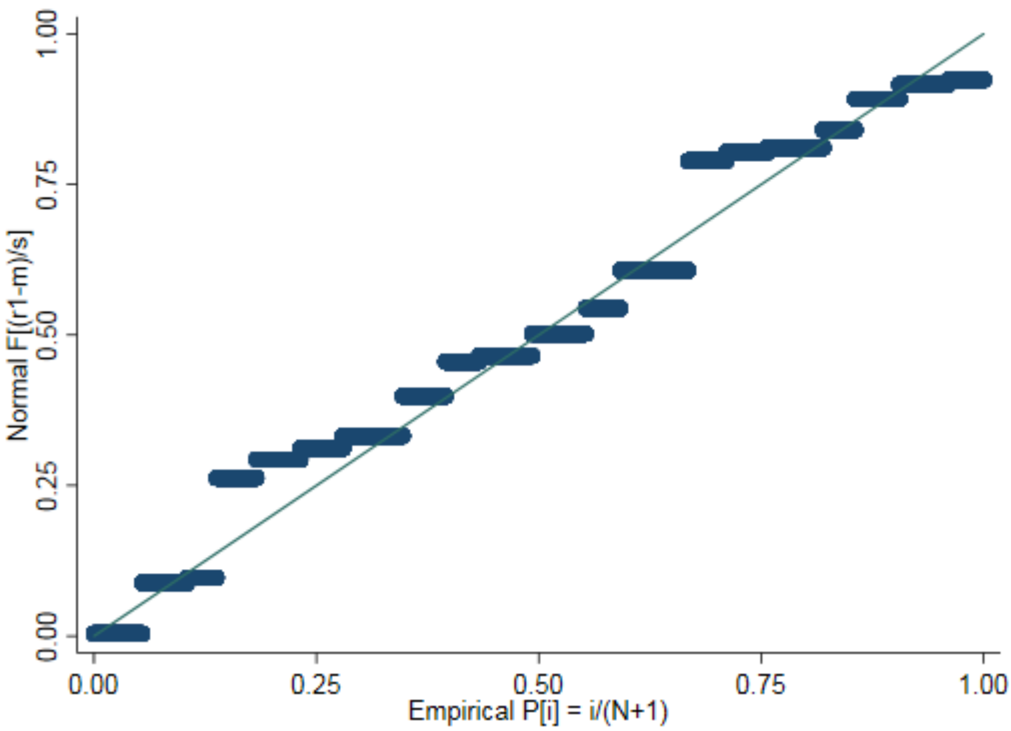
Check the normality of the level-1 residuals via a normal plot.
First calculate the level 1 residuals:

predict lev1, resid
kdensity lev1, normal
pnorm lev1
qnorm lev1

The level 1 residuals lie very close to the straight line and it appears that the assumption of normality for the level 1 residuals is valid.

Repeat the same process for the level 2 residuals:

predict lev2, reffects
kdensity lev2, normal
pnorm lev2
qnorm lev2



The level 2 residuals do not lie perfectly on the diagonal line, but appear to be reasonably close to it so that the assumption of normality seems fair.

12) Examine the residuals of the country effects.
If needed run the random intercept model again with centage as explanatory variable.

mixed hed centage|| country: , ml variance

Now we need to estimate the country-level residuals ($\hat{u}_{oj}$) and their standard errors.
predict u0, reffects
predict u0se, reses

Only choose one pair per each combination of countries. Summary statistics and graphs for country-level variables must be limited on a single record per country. Use the *egen* command with the option *tag* to generate a dummy variable that picks one single observation per country.

egen pickone = tag(country)
sort country

Sort cases by their residual in ascending order.

sort u0

The next step is to rank all the countries in the dataset. The *sum* function creates a new variable equal to the cumulative sum of the variable 'pickone'.

generate u0rank = sum(pickone)

sort country

List will return a list with all the estimates for countries.

list country u0 u0se u0rank if pickone==1

The *serrbar* will create a 'caterpillar plot' to show the country effects in rank order together with 95% confidence intervals. The order of the three variables that follow the command is important. The first variable must contain the point estimates, the second the associated standard errors and the third the rank of the point estimates. We use the scale(1.96) option to obtain 95% confidence limits and the yline(0) option to plot a horizontal line at zero which represents the average school in the data:

serrbar u0 u0se u0rank if pickone==1, scale(1.96) yline(0) saving(P512a, replace)