

**PRI WORKSHOP – Hierarchical Linear Modeling with STATA**

**Vivien W. Chen**  
**Pennsylvania State University**  
**812 Oswald**  
**[vchen@pop.psu.edu](mailto:vchen@pop.psu.edu)**

**12-1pm**  
**2006/12/14**

**Start with basic idea of variance components:**

$$\text{Var}(y_{ij}) = \text{Var}(\zeta_j + \varepsilon_{ij}) = \varphi + \theta$$

Meaning: Total variance = between-subject variances + within-subject variances

Proportion of total variance due to subjects:

$$\rho = \varphi / (\varphi + \theta)$$

In STATA outputs,

sigma\_u = square root of  $\varphi$

-> random intercept of level-2

sigma\_e = square root of  $\theta$

-> within-subject standard deviation

\_cons = overall mean  $\beta$

## I. Random-intercept models:

Example data: <http://www.stata-press.com/data/mlmus/neighborhood.dta>

### Variables:

Level-1: students

attain: educational attainment  
p7vrq: verbal test quotient  
p7read: reading test score  
dadocc: father's occupation  
dadunemp: father is unemployed (dummy)  
momed: mother's education  
male: student is male (dummy)

Level-2: neighborhoods

neighed: neighborhood identifier  
deprive: social-deprivation score

### Random-intercept model

(1) **xtreg** attain p7vrq p7read dadocc male, **i**( neighid) **mle**

```
/* Statement note:  
i: identifier to specify different level of subjects  
mle: maximum likelihood estimation  
*/
```

(2) **xtmixed** attain p7vrq p7read dadocc male || neighid:, **mle**

```
/* Statement note:  
|| : separate random part from fixed part of this model. On the right side is the  
random part.  
*/
```

(3) **gllamm** attain p7vrq p7read dadocc male,**i**(neighid)

```
/* not recommend to use for this model because it takes longer time to get the  
results.*/
```





Result (3)

**. gllamm attain p7vrq p7read dadocc male,i(neighid)**

number of level 1 units = 2310  
number of level 2 units = 524

Condition Number = 39.335539

If condition number is too large, the model may not be well identified.

gllamm model

log likelihood = -2421.5495

attain	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
p7vrq	.0288787	.0022846	12.64	0.000	.024401	.0333564
p7read	.0277196	.001762	15.73	0.000	.0242662	.031173
dadocc	.0117817	.0013151	8.96	0.000	.0092042	.0143592
male	-.0502482	.0288026	-1.74	0.081	-.1067002	.0062039
_cons	.1092556	.0209149	5.22	0.000	.0682632	.1502481

Variance at level 1

.46156532 (.01507852)

Variances and covariances of random effects

\*\*\*level 2 (neighid)

var(1): .01618637 (.00769106)

**Note:**

Compare the results in page 3-5, xtreg and xtmixed have the same results whether in fixed parts or random parts; however, gllamm output has different random part. Usually, the level-1 & 2 variances in gllamm output are the square of the estimates of random-effects parameters computed by using xtmixed and xtreg. If this is not the case, one should improve accuracy by increasing the number of integration points using options **nip()** and **adapt**.

Between-subject effect ( regression on group means)

**. xtreg attain p7vrq p7read dadocc male,i(neighid) be**

```

Between regression (regression on group means)  Number of obs   =   2310
Group variable (i): neighid                    Number of groups =   524

R-sq:  within = 0.4220                      Obs per group:  min =   1
        between = 0.6023                      avg =   4.4
        overall = 0.5175                      max =   16

sd(u_i + avg(e_i.))= .4496683                F(4,519)         =   196.53
                                                Prob > F         =   0.0000
    
```

attain	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
p7vrq	.039306	.005065	7.76	0.000	.0293555	.0492564
p7read	.0215926	.0037577	5.75	0.000	.0142105	.0289747
dadocc	.0172296	.0028122	6.13	0.000	.0117049	.0227542
male	.0339733	.0674072	0.50	0.614	-.0984513	.1663979
_cons	.0736763	.038618	1.91	0.057	-.0021905	.1495431

/\* Statement note:  
be: between effect\*/

Within-subject effect (Fixed-effects model)

**. xtreg attain p7vrq p7read dadocc male,i(neighid) fe**

```

Fixed-effects (within) regression              Number of obs   =   2310
Group variable (i): neighid                    Number of groups =   524

R-sq:  within = 0.4351                      Obs per group:  min =   1
        between = 0.5916                      avg =   4.4
        overall = 0.5225                      max =   16

corr(u_i, Xb) = 0.1797                        F(4,1782)       =   343.11
                                                Prob > F         =   0.0000
    
```

attain	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
p7vrq	.0262106	.0025594	10.24	0.000	.0211909	.0312303
p7read	.0268791	.0020057	13.40	0.000	.0229453	.0308129
dadocc	.0078984	.0015249	5.18	0.000	.0049076	.0108893
male	-.0564366	.0317744	-1.78	0.076	-.1187556	.0058824
_cons	.1120921	.0208618	5.37	0.000	.071176	.1530082
sigma_u	.46132487					
sigma_e	.67411917					
rho	.31894838	(fraction of variance due to u_i)				

F test that all u\_i=0: F(523, 1782) = 1.24 Prob > F = 0.0011

/\* Statement note  
fe: fixed effect \*/

## II. Multilevel Models

### 2-levels variance- component model

Data source: <http://www.stata-press.com/data/mlmus/hsb.dta>

#### **. xtmixed mathach || schoolid:, mle**

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -23557.905

Iteration 1: log likelihood = -23557.905

Computing standard errors:

Mixed-effects ML regression	Number of obs	=	7185
Group variable: schoolid	Number of groups	=	160
	Obs per group: min	=	14
	avg	=	44.9
	max	=	67

Log likelihood = -23557.905	Wald chi2(0)	=	.
	Prob > chi2	=	.

mathach	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----					
_cons	12.63707	.2436173	51.87	0.000	12.15959 13.11455
-----+-----					

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
-----+-----			
schoolid: Identity			
sd(_cons)	2.924632	.1826955	2.587608 3.305552
-----+-----			
sd(Residual)	6.256868	.0527937	6.154245 6.361202
-----+-----			

LR test vs. linear regression: chibar2(01) = 983.92 Prob >= chibar2 = 0.0000

**.xtmixed mathach sector || schoolid:, mle**

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -23539.553  
Iteration 1: log likelihood = -23539.553

Computing standard errors:

Mixed-effects ML regression  
Group variable: schoolid

Number of obs	=	7185
Number of groups	=	160
Obs per group: min	=	14
avg	=	44.9
max	=	67

Log likelihood = -23539.553

Wald chi2(1)	=	41.34
Prob > chi2	=	0.0000

mathach	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
sector	2.804807	.436227	6.43	0.000	1.949817 3.659796
_cons	11.39306	.2909744	39.15	0.000	10.82276 11.96336

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]
schoolid: Identity			
sd(_cons)	2.565071	.1655055	2.260359 2.91086
sd(Residual)	6.257128	.0528	6.154493 6.361475

LR test vs. linear regression: chibar2(01) = 715.25 Prob >= chibar2 = 0.0000

.

### III. Growth-Curve Model

Two ways to fit this model are by using **xtmixed** or **gllamm**.

Use xtmixed:

Quadratic growth model with random intercept:

```
xtmixed y x x2 || id:, mle
```

Quadratic growth model with random intercept and random slope:

```
xtmixed y x x2 || id: age, cov(unstr) mle
```

/\* Statement note:

age: the covariate that we want to obtain the slope

cov():specifies the structure of the (co)variance matrix

for the random effects and may be specified for each  
random-effects equation

unstr: unstructured covariances. This option allows all variances  
and covariances to be distinct. If an equation consists of p  
random effects, the unstructured covariance matrix will  
have  $p(p+1)/2$  parameters to be estimated.

Use gllamm:

Quadratic growth model with random intercept:

```
gen cons=1
```

```
eq inter:cons
```

```
gllamm y x x2, i(id) eqs(inter) adapt
```

Quadratic growth model with random intercept and random slope:

```
matrix a=e(b)
```

```
eq slope: age
```

```
gllamm y x x2, i(id) nrf(2) eqs(inter slope) ip (m) nip(15) from(a)  
adapt
```

/\* For detailed statement, please see Appendix A issued at workshop, or  
get help manual from STATA by typing **h gllamm**. \*/

## Other useful statements & models

(1) Describe the participation pattern in the dataset:

**xtdes, i(subject) t(time)**

/\* statements:

xtdes: describe the different level of participation

i(): identifier

t(): time series variable

\*/

(2) Obtain the overall and within standard deviation:

**xtsum var, i(subject)**

/\* statements:

xtsum: summarize overall and within standard deviation

i(): identifier

\*/

(3) Check on residuals:

To get the distribution of level-1 residuals and level-2 standard deviation.

- To obtain level-1 residual  
**gllapred lv1, pearson**
- To obtain level-2 standard deviation  
**gllapred lv2, ustd**

/\* You have to run gllamm models before use gllapred statement. Alternatively, you can use predict, preserve, then restore; or estimate, then restore, to save residuals and standard deviation. \*/

To obtain posterior means of the predicted probabilities:

**gllapred mu, mu**

Create a histogram for level 2 standard deviation and add a normal curve.

**histogram lv2 if time= =1,normal**

(4) **3-level logistic model**

**gllamm y x1-xn, family(binomial) link(logit) i(level2id level3id)**

**gllamm, eform**

## References

Rabe-Hesketh, Sophia and Anders Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata*. College Station, Texas: Stata Press

GLLAMM (<http://www.gllamm.org/>)

UCLA Stat Computing Portal (<http://statcomp.ats.ucla.edu/mlm/default.htm>)