Lab 5: GROWTH CURVE MODELING

(from pages 78-87 and 91-94 of the old textbook edition and starting on page 210 of the new edition)

Data: Weight gain in Asian children in Britain.

Variables

- **id**: child identifier
- **weight**: weight in Kg
- **age**: age in years
- **gender**: child's gender (1: male, 2: female)

Goal: Use `xtmixed` and `gllamm` to investigate how children grow as they age

```
. use http://www.stata-press.com/data/mlmus/asian, clear
. label def g 1 "boy" 2 "girl"
. label values gender g
```

Exploratory Data Analysis

How many children do we have in the study and how many times did they have their weight measured?

Note that we have to generate a time variable because in order to use the `xtdes` command, STATA needs the time variable to be an integer and age is reported in (non-integer) years.

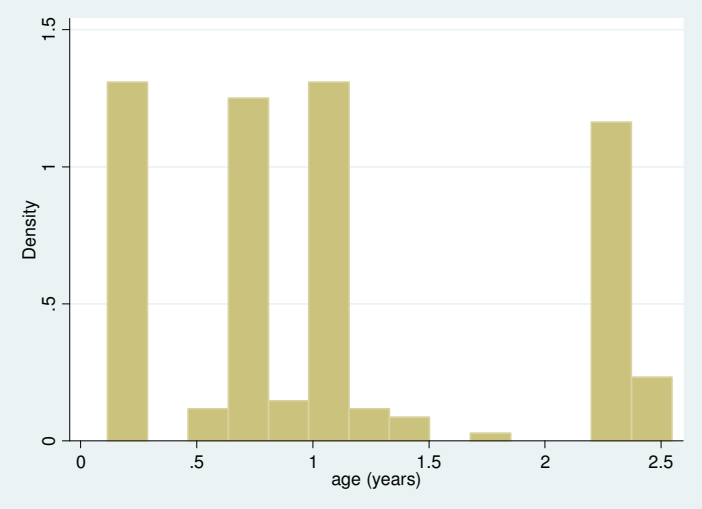
```
. by id: gen time=_n
. xtset id time
       panel variable:  id (unbalanced)
        time variable:  time, 1 to 5
                delta:  1 unit
. xtdes

      id:  45, 258, ..., 4975                                 n =         68
    time:  1, 2, ..., 5                                       T =          5
           Delta(time) = 1; (5-1)+1 = 5
           (id*time uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         3         4       4       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
       27     39.71   39.71 |  111..
       19     27.94   67.65 |  11...
       15     22.06   89.71 |  1111.
        4      5.88   95.59 |  1....
        3      4.41  100.00 |  11111
 ---------------------------+---------
       68    100.00         |  XXXXX
```

We have 68 children, with a maximum of 5 observations per child (3 children) and minimum of 1 observation per child (4 children). The most common number of observations per child (the mode) is 3, since 27 children have 3 observations.

For the analysis, we'll be looking at how weight changes as the children age. It is important to understand the typical ages at which the children have their weight measured.

```
. sum age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       198    1.080552     .787069   .1149897   2.546201
```

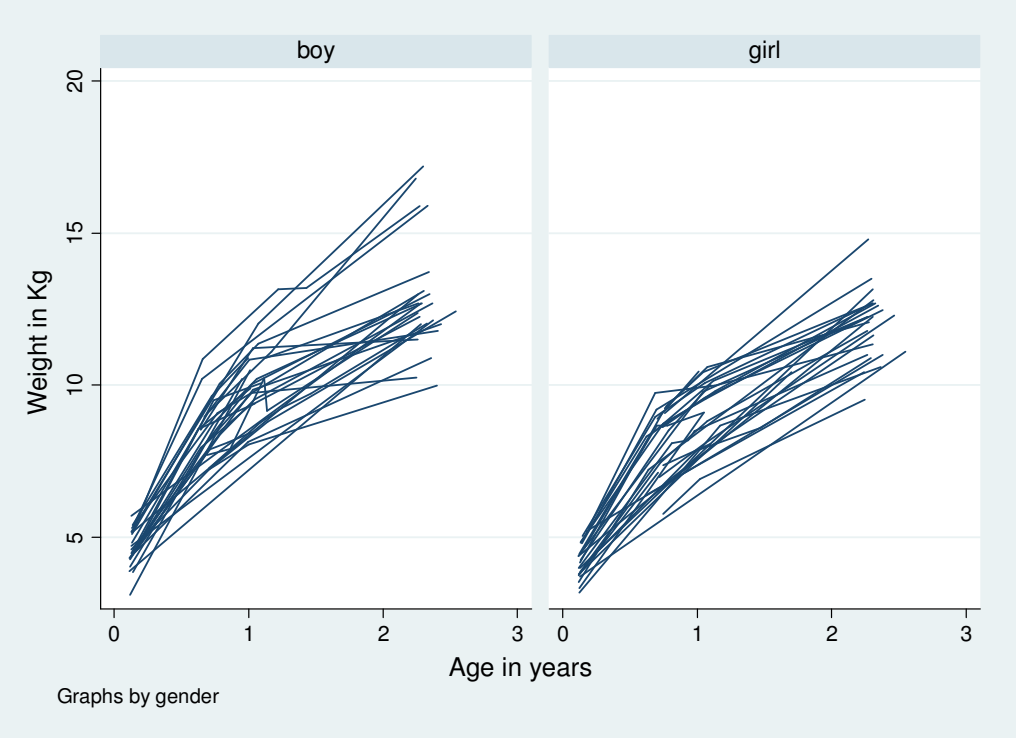
```
. hist age, xtitle(age (years))
```



Weights are generally measured on children at ages 0.2 years (10 weeks), and at 0.7 years (8 months), 1 year, and 2.25 years (27 months)

Now let's take a look how weight changes over time for each child, separately for boys and girls.

```
. sort id age

. graph twoway (line weight age, connect(ascending)), by(gender)
xtitle(Age in years) ytitle(Weight in Kg)
```



What kind of model should we build?
The childrens' growth appears to be non-linear in relation to time. Both boys and girls grow more quickly at first and then they continue to grow, but at a slower rate. Since the relationship between weight and age is non-linear, we will include a **quadratic term for age** in our model. Note that at the first weight measurement, it appears that each child has his or her own starting weight and tends to be at the same weight ranking compared to the other children throughout his or her growth trajectory. We could consider these starting weights to be an approximately normally distributed random variable (around the mean starting weight). We will build a **random intercept** into our initial model.

`xtmixed`

Quadratic growth with random intercept model where U_{1i} is the random intercept for child *i*:

$$weight_{ij} \mid age_{ij}, U_{1i} = \beta_1 + \beta_2 age_{ij} + \beta_3 age_{ij}^2 + U_{1i} + \varepsilon_{ij}$$
$$= (\beta_1 + U_{1i}) + \beta_2 age_{ij} + \beta_3 age_{ij}^2 + \varepsilon_{ij}$$

$$U_{1i} \sim N(0, \tau^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

```
. ** quadratic growth with random intercept **

. gen age2 = age^2
. xtmixed weight age age2 || id:, mle

Mixed-effects ML regression                     Number of obs      =       198
Group variable: id                              Number of groups   =        68

                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5


                                                Wald chi2(2)       =   2623.63
Log likelihood = -276.83266                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.817918   .2896529    26.99   0.000     7.250209    8.385627
        age2 |  -1.705599   .1085984   -15.71   0.000    -1.918448    -1.49275
       _cons |   3.432859   .1810702    18.96   0.000     3.077968     3.78775
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   .9182256   .0973788      .7458965    1.130369
-----------------------------+------------------------------------------------
                sd(Residual) |   .7347063   .0452564      .6511507    .8289837
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    78.07 Prob >= chibar2 = 0.0000
```

Both of the age terms in our model are statistically significant. (If the age^2 term had not been statistically significant we could have included only a linear term for age in our model.) The estimated standard deviation of the random intercept is 0.918 and the estimated standard deviation of the error is 0.734.

Quadratic growth with random intercept U_{1i} and random slope U_{2i} for child *i*:

$$weight_{ij} \mid age_{ij}, U_{1i}, U_{2i} = \beta_1 + \beta_2 age_{ij} + \beta_3 age_{ij}^2 + U_{1i} + U_{2i} age_{ij} + \varepsilon_{ij}$$

$$= (\beta_1 + U_{1i}) + (\beta_2 + U_{2i}) age_{ij} + \beta_3 age_{ij}^2 + \varepsilon_{ij}$$

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix} \right)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

By including a random slope on age, we allow children to have different overall rates of growth.

```
. ** quadratic growth with random intercept and random slope **

. xtmixed weight age age2 || id: age, cov(unstr) mle

Mixed-effects ML regression                     Number of obs      =       198
Group variable: id                              Number of groups   =        68

                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5


                                                Wald chi2(2)       =   1978.20
Log likelihood = -258.07784                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.703998   .2394082    32.18   0.000     7.234767    8.173229
        age2 |  -1.660465   .0885229   -18.76   0.000    -1.833967   -1.486963
       _cons |   3.494512   .1372636    25.46   0.000      3.22548    3.763544
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                     sd(age) |   .5040802   .0879337       .358107    .7095558
                   sd(_cons) |   .6359558   .1293523      .4268684    .9474578
              corr(age,_cons) |   .2747814   .3309063     -.3965135    .7546038
-----------------------------+------------------------------------------------
                sd(Residual) |   .5757751   .0505985      .4846745    .6839993
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =   115.58   Prob > chi2 = 0.0000

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

The standard deviation of the random coefficient on age is 0.50 (95% CI: 0.358, 0.710) which doesn't include 0, so we have evidence that there is heterogeneity between children in growth rates. Also, the estimated standard deviation of the error term has decreased from 0.73 to 0.57 indicating better fit of the model.

What if there is a systematic different in growth between boys and girls?

Quadratic growth with random intercept U_{1i} and random slope U_{2i} for child *i* that includes a child-level covariate, an indicator of gender:

$$weight_{ij} \mid age_{ij}, girl_{ij}, U_{1i}, U_{2i} = (\beta_1 + U_{1i}) + (\beta_2 + U_{2i}) age_{ij} + \beta_3 age_{ij}^2 + \beta_4 girl_{ij} + \varepsilon_{ij}$$

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix} \right)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

```
. ** including a child-level covariate **
. gen girl = gender - 1

. xtmixed weight age age2 girl || id: age , cov(unstr) mle

Mixed-effects ML regression                     Number of obs      =       198
Group variable: id                              Number of groups   =        68

                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5


                                                Wald chi2(3)       =   1975.44
Log likelihood = -253.86692                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.697967   .2382121    32.32   0.000      7.23108    8.164855
        age2 |  -1.657843   .0880529   -18.83   0.000    -1.830423   -1.485262
        girl |  -.5960093   .1963689    -3.04   0.002    -.9808853   -.2111332
       _cons |   3.794769   .1655053    22.93   0.000     3.470385    4.119153
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                     sd(age) |   .5097089   .0871791      .3645317    .7127039
                   sd(_cons) |    .594731   .1289891      .3887823    .9097762
              corr(age,_cons) |   .1571086   .3240801     -.4564674    .6694143
-----------------------------+------------------------------------------------
                sd(Residual) |   .5723301   .0496274      .4828786    .6783521
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =   104.17   Prob > chi2 = 0.0000

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

Interpretation of the coefficient on girl:
At any given age, we estimate that a boy is 0.60 kg heavier than he would be if he were a girl. In other words, at any given age, we estimate that the typical (U_{1i}=0) boy will be 0.60 kg heavier than the typical (U_{1i}=0) girl.

gllamm

When modeling random effects (beyond a random intercept) in gllamm, we need to use the **eq** command to specify the 'equation' for each random effect. The 'equation' is the variable or constant by which we multiply the random effect. For example, if we were creating the equation for a random intercept we would multiply the random effect by 1. If we were creating the equation for a random coefficient on the variable x we would multiply the random effect by variable x.

We include the name of each equation in the **eqs** option of gllamm.

```
. ** quadratic growth with random intercept **

. gen cons = 1
. eq inter: cons
. gllamm weight age age2, i(id) eqs(inter) adapt


number of level 1 units = 198
number of level 2 units = 68


gllamm model

log likelihood = -276.83266

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.817871   .2899873    26.96   0.000     7.249507    8.386236
        age2 |  -1.705589   .1086957   -15.69   0.000    -1.918629    -1.49255
       _cons |   3.432893   .1811779    18.95   0.000      3.07779    3.787995
------------------------------------------------------------------------------

Variance at level 1
------------------------------------------------------------------------------

  .53966034 (.06647545)

Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (id)

    var(1): .84334423 (.17887769)
------------------------------------------------------------------------------


. ** quadratic growth with random intercept and random slope **

. eq slope: age
```

The option **`nrf(2)`** specifies that we now have two random effects (intercept and slope). The `ip(m) nip(15)` specifies that we are using a spherical integration rule of degree 15 (don't need to worry about this – just know that it speeds up the estimation).

```
. gllamm weight age age2, i(id) nrf(2) eqs(inter slope) ip(m) nip(15) adapt

number of level 1 units = 198
number of level 2 units = 68

Condition Number = 8.9386847

gllamm model

log likelihood = -258.07784

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   7.703998     .24026    32.07   0.000     7.233097    8.174899
        age2 |  -1.660465   .0890109   -18.65   0.000    -1.834923   -1.486007
       _cons |   3.494512   .1376254    25.39   0.000     3.224771    3.764253
------------------------------------------------------------------------------
```

```
Variance at level 1
------------------------------------------------------------------------------

  .3315169 (.05826676)

Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (id)

    var(1): .40444011 (.16452483)
    cov(2,1): .0880873 (.08802562) cor(2,1): .27478078

    var(2): .25409706 (.08865135)
------------------------------------------------------------------------------
```

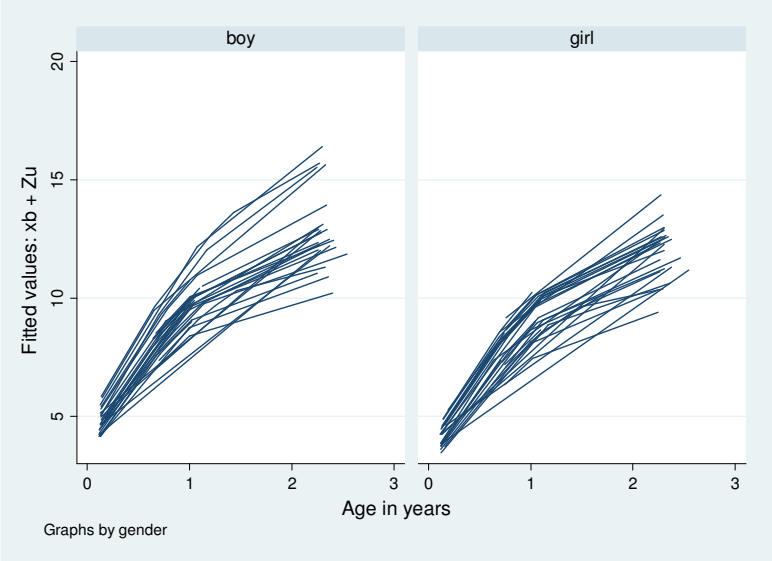
If you compare the results from xtmixed and gllamm, you'll see that they are similar. You can get the gllamm results to be even closer to the results from xtmixed if you increase `nip()`.

Predicting the trajectories for each child

- **`xtmixed`**

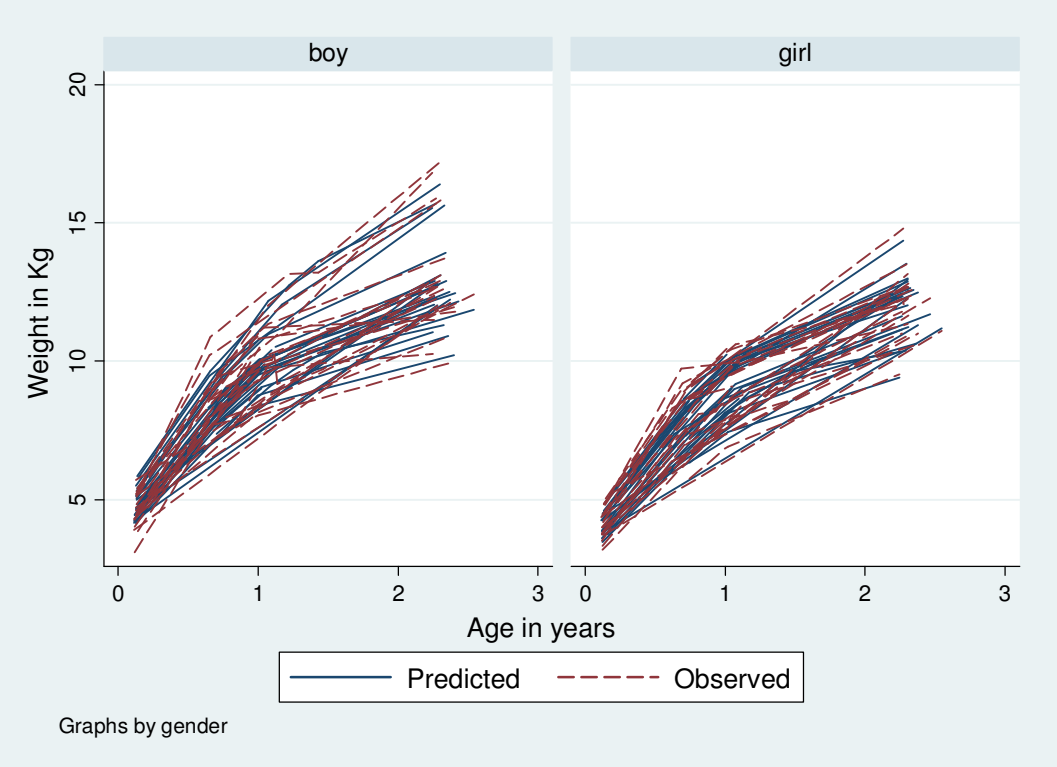
Get the empirical Bayes estimates of the random intercepts and random slopes

```
. * re-run the xtmixed including the child-level covariate
. xtmixed weight age age2 girl || id: age , cov(unstr) mle

. predict traj, fitted

. sort id age

* plot only the predicted *
. graph twoway (line traj age, connect(ascending)), by(gender)
xtitle(Age in years)
```



In the above plot you can see how including a quadratic term for age allowed the relationship between age and predicted weight to be nonlinear since the trajectories are not straight lines and tend to have a steeper rise at earlier ages and then rise more slowly at older ages. The fixed effect for gender allowed for a systematic difference in predicted weight for boys and girls. The trajectories for boys tend to be higher than for girls of the same age. The random intercept is reflected in the different 'starting point' for each of the trajectories. We also see the random coefficient on age reflected by the different rates of growth for different children.

```
* plot the predicted and the observed *
. graph twoway (line traj age, connect(ascending)) (line weight age,
connect(ascending) clpatt(dash)), by(gender) xtitle(Age in years)
ytitle(Weight in Kg) legend(order(1 "Predicted" 2 "Observed"))
```



The model appears to fit the data adequately based on a comparison of the fitted trajectories to the observed trajectories.

- **`gllamm`**

Get the empirical Bayes estimates of the random intercepts and random slopes

```
. * re-run the gllamm including the child-level covariate
. gllamm weight age age2, i(id) nrf(2) eqs(inter slope) ip(m) nip(15)
adapt

. gllapred traj, linpred

. graph twoway (line traj age, connect(ascending)) (line weight age,
connect(ascending) clpatt(dash)), by(gender) xtitle(Age in years)
ytitle(Weight in Kg) legend(order(1 "Predicted" 2 "Observed"))
```

This will produce a similar graph to the one we saw for `xtmixed`.

I'm not convinced that the third model (including a random intercept, random slope on age and a fixed effect for girl) is the 'best' of the three models. Are there model selection criteria I can use?

Yes! You can look at AIC and BIC for either `xtmixed` or `gllamm`. According to the AIC or BIC criterion, the best fitting model is the model that has the smallest value of AIC or BIC, respectively.

```
. ** quadratic growth with random intercept **
. quietly xtmixed weight age age2 || id:, mle

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    198           .   -276.8327      5     563.6653    580.1067
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

. ** quadratic growth with random intercept and random slope **
. quietly xtmixed weight age age2 || id: age, cov(unstr) mle

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    198           .   -258.0778      7     530.1557    553.1736
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

. ** including a child-level covariate **
. quietly xtmixed weight age age2 girl || id: age , cov(unstr) mle

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    198           .   -253.8669      8     523.7338      550.04
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

The third models wins in terms of both AIC and BIC.

The code to do the same thing is gllamm is very similar (check out the .do file for this lab).