Lab 4: Two-level Random Intercept Model

Data: Peak expiratory flow rate (pefr) measured twice, using two different instruments, for 17 subjects. (from Chapter 1 of *Multilevel and Longitudinal Modeling Using Stata*)

Goals:
1. Review how to fit a random intercept model using xtreg, xtmixed and gllamm.
2. Interpret parameters in a random intercept model.
3. Model measurement error with random intercept model.
4. Obtain predictions from multilevel model.

PART I Exploratory Data Analysis

Data Structure:

```
     +----------------------------+
     | id   wp1   wp2   wm1   wm2 |
     |----------------------------|
  1. |  1   494   490   512   525 |
  2. |  2   395   397   430   415 |
  3. |  3   516   512   520   508 |
  4. |  4   434   401   428   444 |
  5. |  5   476   470   500   500 |
```

Variables
- id: subject id
- wp1: Wright peak, occasion 1
- wp2: Wright peak, occasion 2
- wm1: Mini Wright, occasion 1
- wm2: Mini Wright, occasion 2

- Dataset is in wide format.
- Repeated measurements of wp and wm are nested within subject.
- No missing data

Exploratory Analysis (We will only work with wm for now):

First, calculate the overall mean lung function and store it as a local variable, `wm_mean`.

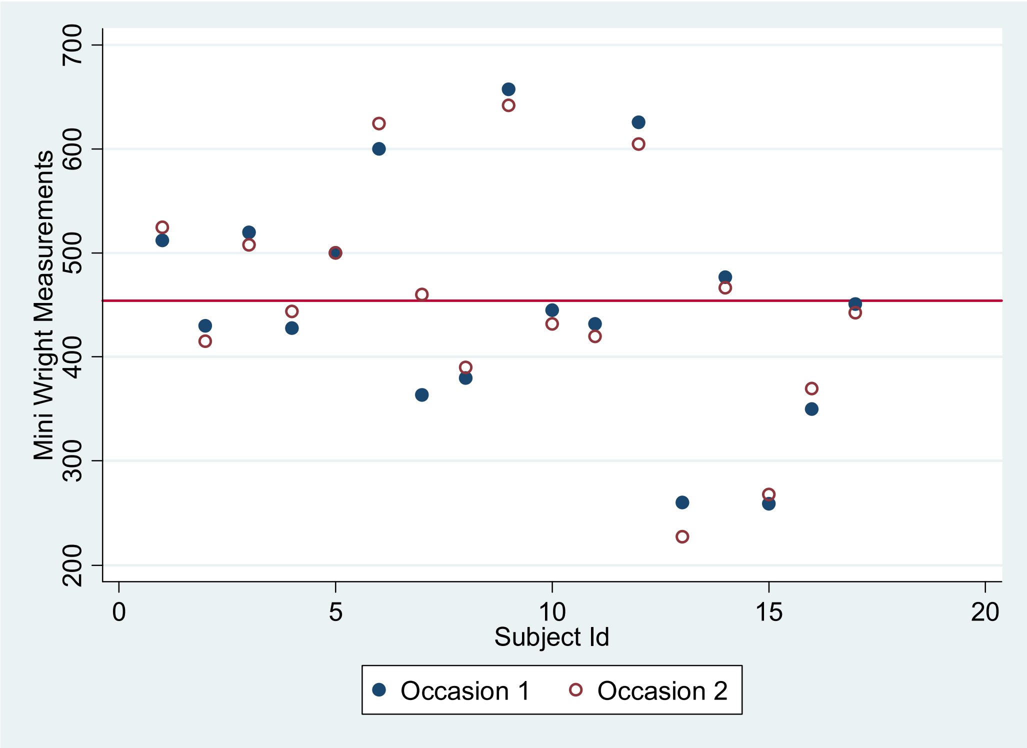
```
. generate mean_wm = (wm1+wm2)/2
. summarize mean_wm

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     mean_wm |        17    453.9118    111.2912      243.5        650

. local wm_mean = r(mean)
```

Let's display the values of the repeated Mini Wright meter measures of lung function for each subject and the overall mean lung function.

```
. twoway (scatter wm1 id, msymbol(circle)) (scatter wm2 id,
symbol(circle_hollow)), xtitle(Subject Id) ytitle(Mini Wright Measurements)
legend(order(1 "Occassion 1" 2 "Occasion 2")) yline(`wm_mean')
```



- Measurements taken from the same person were clustered together.

- It appears that the meann of the two observations for each individual are normally scattered (like a normal distribution) around the overall mean.

Might this suggest a subject-level random intercept model?

(1) For an individual *i*, the two repeated Mini Wright values (y_{i1} and y_{i2}) are trying to capture the same *true* peak expiratory flow rate (β_i) that is unobservable.

(2) Let's assume what we actually measured is the true value (β_i) plus some random (measurement) error (ε_{ij}). So

$$y_{ij} = \beta_i + \varepsilon_{ij}$$

(3) Note that this looks like our typical random-intercept model:

$$y_{ij} = \beta + v_i + \varepsilon_{ij}$$

where $\beta_i = \beta + v_i$.
By writing β_i this way, we also allow this model to accommodate pefr from *different* people.

(4) Now let's include the random components of our model:

A measurement error distribution that is identical for each individual:

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

A distribution describing the variation in the true pefr in the population:

$$v_i \sim Normal(0, \tau^2)$$

(5) Our final model:

$$y_{ij} = \beta + v_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim Normal(0, \sigma^2), \qquad v_i \sim Normal(0, \tau^2)$$

Note that here β can be interpreted as the average true pefr in the population (similar to the red line in the above graph). How would you describe the other model parameters' presence in the scatter plot above?

Reshape Data

We need to reshape the data to a 'long' format for the data analysis.

```
. reshape long wm wp, i(id) j(occasion)
note: j = 1 2)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       17   ->     34
Number of variables                   5   ->      4
j variable (2 values)                     ->   occasion
xij variables:
                                wm1 wm2   ->   wm
                                wp1 wp2   ->   wp
-----------------------------------------------------------------------------

     +---------------------+
     | id   occas~n     wm |
     |---------------------|
  1. |  1         1    512 |  (i = 1, j = 1)
  2. |  1         2    525 |  (i = 1, j = 2)
  3. |  2         1    430 |  (i = 2, j = 1)
  4. |  2         2    415 |  (i = 2, j = 1)
  5. |  3         1    520 |
```

More Exploratory Analysis:

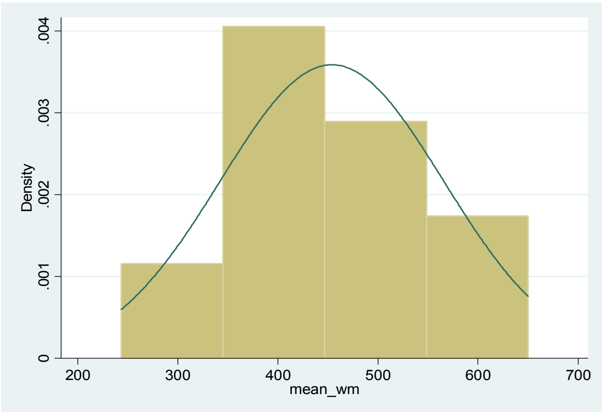
Let's check some of the distributional assumptions (note that we only have 17 people).

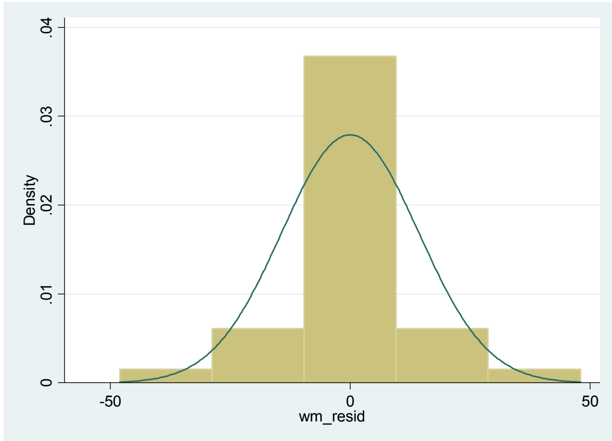
(1) Check $v_i \sim Normal(0, \tau^2)$:

```
sort(id)
by id, egen mean_wm mean(wm)
hist mean_wm, norm
```

(2) Check $\varepsilon_{ij} \sim Normal\left(0, \sigma^2\right)$

```
gen wm_resid = wm-mean_wm
hist wm_resid, norm
```





PART II Fitting the Model and Interpretation

Fitting the random intercept model with "xtreg"

```
. xtreg wm, i(id) mle

Iteration 0:   log likelihood = -187.89003
Iteration 1:   log likelihood = -184.95979
Iteration 2:   log likelihood = -184.76189
Iteration 3:   log likelihood =  -184.5855
Iteration 4:   log likelihood =  -184.5784
Iteration 5:   log likelihood = -184.57839

Random-effects ML regression                    Number of obs      =        34
Group variable (i): id                          Number of groups   =        17

Random effects u_i ~ Gaussian                   Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2

                                                Wald chi2(0)       =      0.00
Log likelihood  = -184.57839                    Prob > chi2        =         .

------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.18616    17.33   0.000     402.5878    505.2357
-------------+----------------------------------------------------------------
    /sigma_u |   107.0464   18.67858                       76.0406    150.6949
    /sigma_e |   19.91083   3.414659                       14.2269     27.8656
         rho |   .9665602   .0159494                      .9210943    .9878545
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=   46.27 Prob>=chibar2 = 0.000
```

- Does the estimate of β (_const) = 453.9118 look familiar?

In the output above, ρ (rho) can be interpreted as either

- the proportion of the total variance that is between subjects (or due to subjects)

$$\rho = \frac{\text{variance.between}}{\text{total.variance}} = \frac{Var(v_i)}{Var(y_{ij})} = \frac{\tau^2}{\tau^2 + \sigma^2}$$

- the **correlation** between the measurements on different occasions for the same subject (intra-class correlation)

$$\rho = Corr(y_{ij}, y_{ij'}) = \frac{Cov(y_{ij}, y_{ij'})}{\sqrt{Var(y_{ij})}\sqrt{Var(y_{ij'})}} = \frac{\tau^2}{\sqrt{\tau^2 + \sigma^2}\sqrt{\tau^2 + \sigma^2}} = \frac{\tau^2}{\tau^2 + \sigma^2}$$

It can be a little confusing because, the ***covariance*** *between measurements on different occasions for the same subject is* σ^2.

Interpretations

- Notice that ρ = .966 is very high! The repeated observations within individuals are highly correlated and the proportion of the total variance that is between subjects is very large.
- `/sigma_u` is 107.05, the estimate of the standard deviation of the random intercepts. Hence we expect about 95% of the random intercepts to fall within 200 (= approximately 107.05*2) units on either direction of the estimated overall mean, 453.91, or in other words, between 250 and 650.
- The estimated within-subject standard deviation is `/sigma_e` = 19.9. Hence we expect 95% of the repeated observations on an individual to fall within 40 (= approximately 19.9*2) units from the subject-specific mean.

The results from xtreg, mle are equivalent to those from xtmixed, mle. The difference between xtreg and xtmixed is that xtreg is designed more for cross-sectional time-series linear regression and can only be used to fit a random intercept. On the other hand, xtmixed is designed for multi-level mixed effects linear regression and can be used to fit random coefficients and different levels of mixed effects.

Fitting the random intercept model with xtmixed

```
. xtmixed wm || id:, mle

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -184.57839
Iteration 1:   log likelihood = -184.57839

Computing standard errors:
```

```
Mixed-effects ML regression                     Number of obs      =        34
Group variable: id                              Number of groups   =        17

                                                Obs per group: min =         2
                                                               avg =       2.0
                                                               max =         2
                                                Wald chi2(0)       =         .
Log likelihood = -184.57839                     Prob > chi2        =         .

------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9118   26.18616    17.33   0.000     402.5878    505.2357
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   107.0464   18.67857       76.0406    150.6949
-----------------------------+------------------------------------------------
                sd(Residual) |   19.91083   3.414679      14.22688    27.86565
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    46.27 Prob >= chibar2 = 0.0000
```

Fitting the random intercept model with gllamm

```
. gllamm wm, i(id) nip(12) adapt

Running adaptive quadrature
Iteration 0:    log likelihood = -207.72022
Iteration 1:    log likelihood = -205.79654
Iteration 2:    log likelihood = -185.72467
Iteration 3:    log likelihood = -184.63453
Iteration 4:    log likelihood = -184.57846
Iteration 5:    log likelihood =  -184.5784
Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood =  -184.5784
Iteration 1:   log likelihood = -184.57839

number of level 1 units = 34
number of level 2 units = 17

Condition Number = 152.64774

gllamm model
log likelihood = -184.57839
------------------------------------------------------------------------------
          wm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   453.9116   26.18394    17.34   0.000      402.592    505.2312
------------------------------------------------------------------------------
Variance at level 1
------------------------------------------------------------------------------
  396.70879 (136.11609)
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (id)
    var(1): 11456.828 (3997.7689)
------------------------------------------------------------------------------
```

Note that gllamm returns variances and not standard deviations.

PART III Prediction

Goal 1: So what is our best estimate of each subject's **true** peak expiratory flow rate

Recall that when constructing our model:

$$y_{ij} = \beta + v_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim Normal(0, \sigma^2), \qquad v_i \sim Normal(0, \tau^2) \quad .$$

So we'd like to obtain the estimated value of $\beta + v_i$ for each individual *i*.
β is given in the output so we need to extract the v_i's.

Estimating the random intercepts using empirical Bayes and gllamm

```
. gllapred eb, u
(means and standard deviations will be stored in ebm1 ebs1)
Non-adaptive log-likelihood: -202.25846
 -245.1480  -225.1857  -211.3252  -199.5193  -190.8173  -186.2250
 -184.7457  -184.5784  -184.5784
log-likelihood:-184.57839
```

Empirical Bayes estimate of the subject-specific mean, i.e. $\beta + v_i$

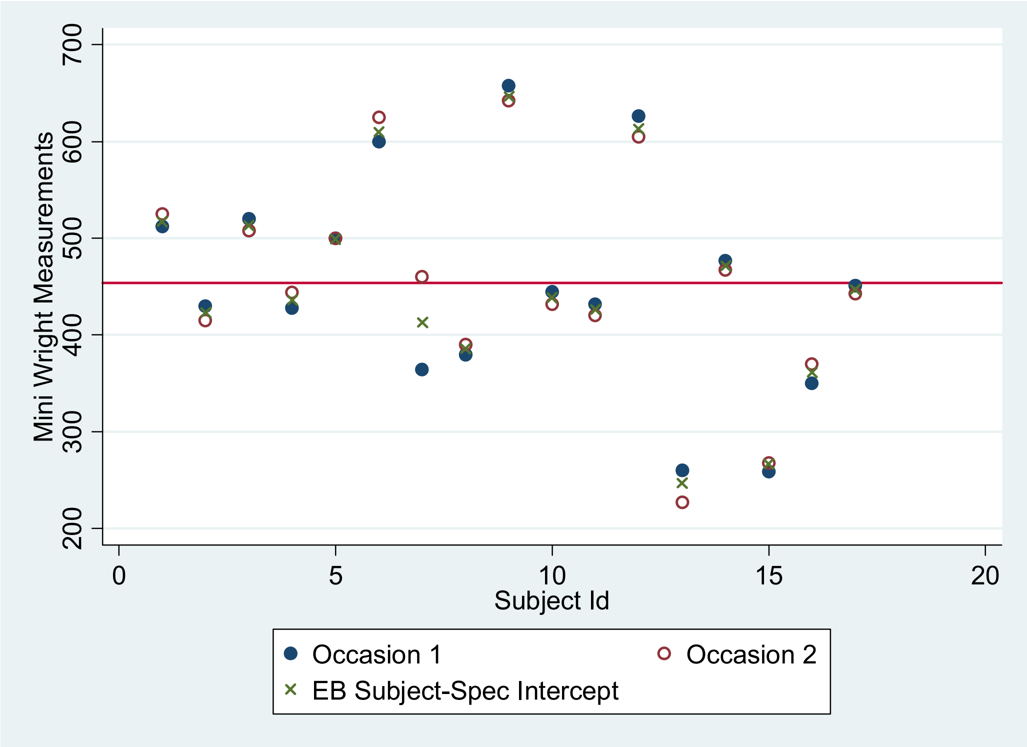
```
. gllapred eb, linpred
(linear predictor will be stored in eb)
Non-adaptive log-likelihood: -202.25846
 -245.1480  -225.1857  -211.3252  -199.5193  -190.8173  -186.2250
 -184.7457  -184.5784  -184.5784
log-likelihood:-184.57839
```

```
. reshape wide wm wp eb ebm1 ebs1, i(id) j(occasion)
(note: j = 1 2)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       34   ->      17
Number of variables                   8   ->      12
j variable (2 values)          occasion   ->   (dropped)
xij variables:
                                     wm   ->   wm1 wm2
                                     wp   ->   wp1 wp2
                                     eb   ->   eb1 eb2
                                   ebm1   ->   ebm11 ebm12
                                   ebs1   ->   ebs11 ebs12
-----------------------------------------------------------------------------
```

Let's plot the estimated peak expiratory flow rate:

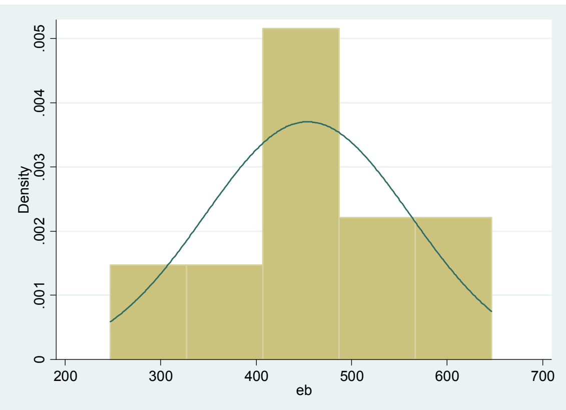
```
. twoway (scatter wm1 id, msymbol(circle)) (scatter wm2 id,
msymbol(circle_hollow)) (scatter eb1 id, msymbol(X)), xtitle(Subject Id)
ytitle(Mini Wright Measurements) legend(order(1 "Occassion 1" 2 "Occasion 2" 3
"EB Subject-Spec Intercept")) yline(`wm_mean')
```

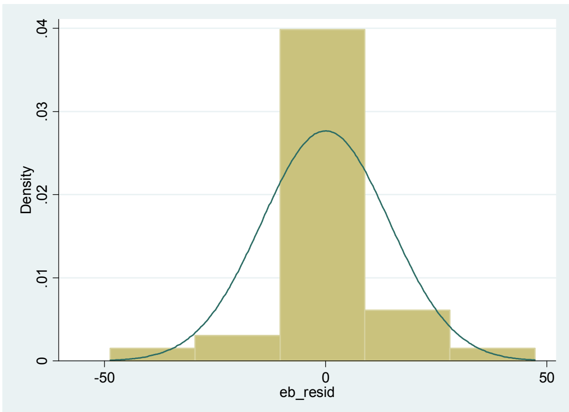


- Note that the estimated peak expiratory flow rate (**x**) do not always fall in between the measurements at occasion 1 and occasion 2!!! Why? (Hint: look at subject 6 and 13).

- Let's check our model assumptions again with the estimated intercepts and residuals:

```
. hist eb, norm
. gen eb_resid = wm-eb
. hist eb_resid, norm
```

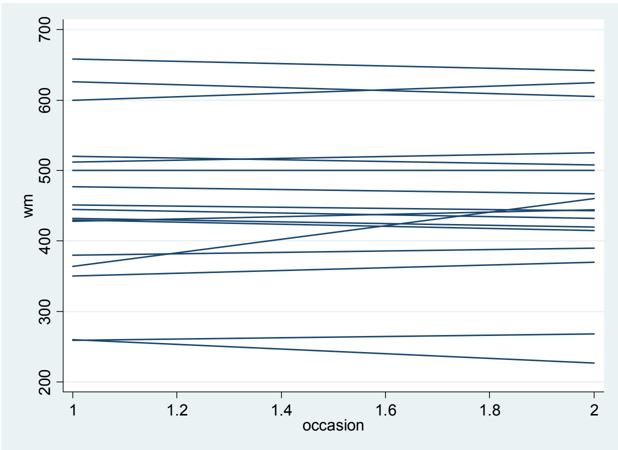




Goal 2: Based on our model, can we make prediction about future observation of a new measurement taken from an existing subject or a new measurement from a new subject?

Extra

- The random effect model above is motivated by measurement error. It's similar to the usual LDA setting where we can view the data as:



- To incorporate both wp and wm measurements in a model we can use a three-level random effect model:
  Subject (level 3) → Method (level 2) → Repeated measurements (level 1)
  See textbook.