BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 11

Linear Mixed Effects Model and Prediction

Linear Mixed Effects Model and Prediction

Recall that in the linear mixed effects model,

$$Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \epsilon_{ij},$$

we can distinguish between the conditional mean,

$$E(Y_{ij}|X_{ij}, b_i) = X_{ij}'\beta + Z_{ij}'b_i,$$

and the marginal mean,

$$E(Y_{ij}|X_{ij}) = X_{ij}'\beta.$$

The former describes the mean response for an individual, the latter describes the mean response averaged over individuals.

The distinction between the conditional and marginal means is best understood with a simple example.

Consider the simple random intercepts and slopes model,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

In this model, we can distinguish the conditional mean for an individual,

$$E(Y_{ij} | b_{1i}, b_{2i}) = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij},$$

(see broken lines for subjects A and B in Figure 1), and the marginal mean averaged over individuals,

$$E(Y_{ij}) = \beta_1 + \beta_2 t_{ij},$$

(see solid line in Figure 1).

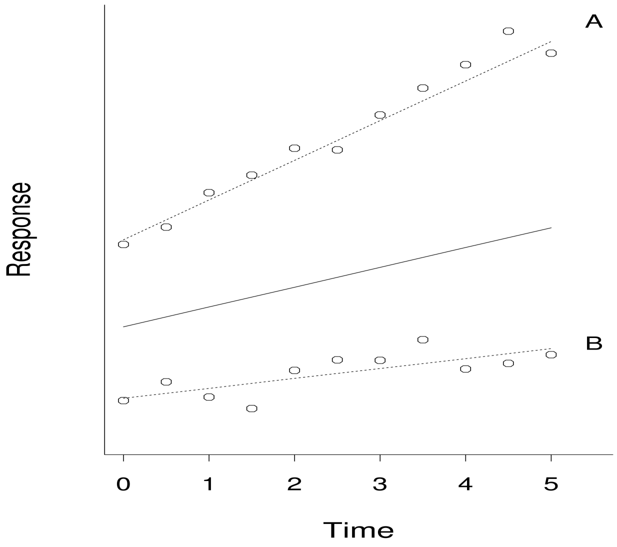

Figure 1: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Prediction of Random Effects

In many applications, inference is focused on fixed effects, $\beta_1, \beta_2, ..., \beta_p$.

However, we can also "estimate" or predict subject-specific effects, b_i (or subject-specific response trajectories over time).

Technically, because the b_i are random, we customarily talk of "predicting" the random effects rather than "estimating" them.

Using maximum likelihood, the prediction of b_i, say $\widehat{b}_i$, is given by:

$$\widehat{b}_i = E(b_i | Y_i; \widehat{\beta}, \widehat{G}, \widehat{\sigma}^2).$$

This is known as "best linear unbiased predictor" (or BLUP).

In general, BLUP "shrinks" predictions towards population-averaged mean.

For example, consider the random intercept model

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij},$$

where $\mathrm{Var}\,(b_i) = \sigma_b^2$ and $\mathrm{Var}\,(\epsilon_{ij}) = \sigma^2$.

It can be shown that the BLUP for b_i is:

$$\widehat{b}_i = w \times \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \mu_{ij}) \right) + (1 - w) \times 0, \text{ where } w = \frac{n_i \sigma_b^2}{n_i \sigma_b^2 + \sigma^2}.$$

That is, a weighted-average of zero (mean of b_i) and the mean "residual" for the i^{th} subject.

Note: Less shrinkage (toward zero) when n_i is large and when σ_b^2 is large relative to σ^2.

For the general case, the prediction of b_i is given by:

$$\widehat{b}_i = E(b_i|Y_i;\widehat{\beta},\widehat{G},\widehat{\sigma}^2) = \widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}),$$

where $\Sigma_i = \text{Cov}\,(Y_i|X_i) = Z_iGZ_i' + R_i = Z_iGZ_i' + \sigma^2 I$.

When the unknown covariance parameters have been replaced by their ML or REML estimates, the resulting predictor is often referred to as the "Empirical BLUP" or the "Empirical Bayes" (EB) estimator.

Finally, the i^{th} subject's predicted response profile is,

$$\begin{aligned}\widehat{Y}_i &= X_i\widehat{\beta} + Z_i\widehat{b}_i \\ &= X_i\widehat{\beta} + Z_i\widehat{G}Z_i'\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta}) \\ &= (\widehat{R}_i\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + (I - \widehat{R}_i\widehat{\Sigma}_i^{-1})Y_i\end{aligned}$$

That is, the i^{th} subject's predicted response profile is a weighted combination of the population-averaged mean response profile, $X_i\widehat{\beta}$, and the i^{th} subject's observed response profile Y_i.

Subject's predicted response profile is "shrunk" towards population-averaged mean response profile.

Amount of "shrinkage" depends on relative magnitude of R_i and Σ_i.

Note that R_i characterizes the within-subject variability, while Σ_i incorporates both within-subject and between-subject sources of variability.

$$R_i\Sigma_i^{-1} = \frac{\text{within-subject variability}}{\text{within-subject + between-subject variability}}.$$

Thus, $R_i\Sigma_i^{-1}$ denotes the fraction of total variability that is due to within-subject (or measurement error) variation.

Similarly, $(I - R_i\Sigma_i^{-1})$ denotes the fraction of total variability that is due to between-subject variation.

When within-subject variability is large relative to between-subject variability, more weight is given to $X_i\widehat{\beta}$, the population-averaged mean response profile (more "shrinkage").

PROC MIXED in SAS

The Empirical Bayes (EB) estimates, $\widehat{b}_i$, can be obtained by using the following option on the RANDOM statement in PROC MIXED:

RANDOM INTERCEPT time / TYPE=UN SUBJECT=id S;

Alternatively, a subject's predicted response profile,

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i,$$

can be obtained by using the following option on the MODEL statement:

MODEL y = trt time trt*time / OUTP=*SAS-data-set***;**

Example: *Exercise Therapy Study*

Consider a model with randomly varying intercepts and slopes, and which allows the mean values of the intercept and slope to differ in the two treatment groups.

To fit this model, use the following SAS code:

```
PROC MIXED DATA=stren;
      CLASS id trt;
      MODEL y=trt time time*trt / S CHISQ;
      RANDOM INTERCEPT time / TYPE=UN SUBJECT=id G S;
```

Empirical Bayes Estimates of b_i

```
                    Solution for Random Effects

                                       Std Err
Effect       id    Estimate       Pred     t Value     Pr > |t|

Intercept     1     -1.0111     0.9621       -1.05       0.2959
time          1    -0.03812    0.08670       -0.37       0.7144
Intercept     2      3.3772     0.9621        1.07       0.0007
time          2      0.1604    0.08670        1.85       0.0672
    .         .         .          .            .            .
    .         .         .          .            .            .
    .         .         .          .            .            .
```

Example: *Exercise Therapy Study*

Next, we consider how to obtain a subject's predicted response profile.

```
PROC MIXED DATA=stren;
     CLASS id trt;
     MODEL y=trt time time*trt / S CHISQ OUTP=predict;
     RANDOM INTERCEPT time / TYPE=UN SUBJECT=id G S;

PROC PRINT DATA=predict;
     VAR id trt time y Pred StdErrPred Resid;
```

Predicted Response Profiles

```
                                  StdErr
id  trt  time   y      Pred       Pred       Resid

 1    1     0  79   78.9937    0.59729     0.00634
 1    1     4  79   79.4071    0.39785    -0.40707
 1    1     6  80   79.6138    0.36807     0.38623
 1    1     8  80   79.8205    0.40451     0.17952
 1    1    12  80   80.2339    0.61057    -0.23389
 2    1     0  83   83.3820    0.59729    -0.38202
 2    1     4  85   84.5644    0.39785     0.43562
 2    1     6  85   85.1556    0.36807    -0.15557
 2    1     8  86   85.7468    0.40451     0.25325
 2    1    12  87   86.9291    0.61057     0.07088
 .    .     .   .      .          .           .
 .    .     .   .      .          .           .
 .    .     .   .      .          .           .
```

Case Study: *Influence of Menarche on Changes in Body Fat*

- Prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study.
- At start of study, all the girls were pre-menarcheal and non-obese
- All girls were followed over time according to a schedule of annual measurements until four years after menarche.
- The final measurement was scheduled on the fourth anniversary of their reported date of menarche.
- At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis.

Consider an analysis of the changes in percent body fat before and after menarche.

For the purposes of these analyses "time" is coded as time since menarche and can be positive or negative.

Note: measurement protocol is the same for all girls.

Study design is almost "balanced" if timing of measurement is defined as time since baseline measurement.

It is inherently unbalanced when timing of measurements is defined as time since a girl experienced menarche.

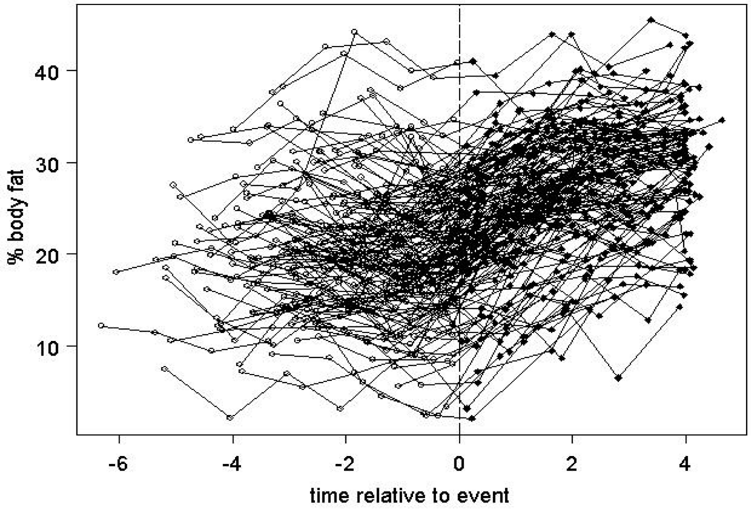

Figure 2: Timeplot of percent body fat against time, relative to age of menarche (in years).

Consider hypothesis that %body fat increases linearly with age, but with different slopes before/after menarche.

We assume that each girl has a piecewise linear spline growth curve with a knot at the time of menarche (see Figure 3).

Each girl's growth curve can be described with an intercept and two slopes, one slope for changes in response before menarche, another slope for changes in response after menarche.

Note: the knot is not at a fixed age for all subjects.

Let t_{ij} denote time of the j^{th} measurement on i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche).

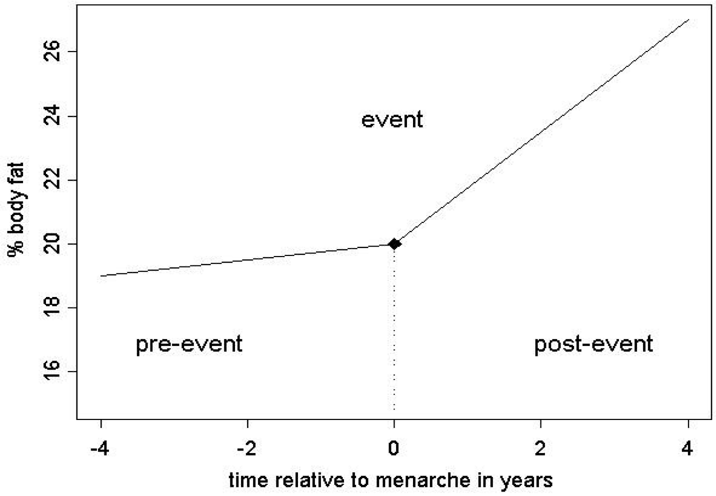

Figure 3: Graphical representation of piecewise linear trajectory.

We consider the following linear mixed effects model

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+,$$

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$.

Interpretation of model parameters:

The intercept β_1 is the average %body fat at menarche (when $t_{ij} = 0$).

The slope β_2 is the average rate of change in %body fat (per year) during the pre-menarcheal period.

The average rate of change in %body fat (per year) during the post-menarcheal period is given by $(\beta_2 + \beta_3)$.

Goal: Assess whether population slopes differ before and after menarche, i.e., $H_0 : \beta_3 = 0$.

Similarly, $(\beta_1 + b_{1i})$ is intercept for i^{th} subject and is the true %body fat at menarche (when $t_{ij} = 0$).

$(\beta_2 + b_{2i})$ is i^{th} subject's slope, or rate of change in %body fat during the pre-menarcheal period.

Finally, the i^{th} subject's slope during the post-menarcheal period is given by $[(\beta_2 + \beta_3) + (b_{2i} + b_{3i})]$.

Interpretation of variance components:

Recall that the subject-specific intercepts, $(\beta_1 + b_{1i})$, have mean β_1 and variance $g_{11} = \sigma^2_{b_{1i}}$.

Furthermore, since $b_{1i} \sim N(0, \sigma^2_{b_{1i}})$ this implies that $(\beta_1 + b_{1i}) \sim N(\beta_1, \sigma^2_{b_{1i}})$.

Under the assumption of normality, we expect 95% of the subject-specific intercepts, $(\beta_1 + b_{1i})$, to lie between: $\beta_1 \pm 1.96 \times \sigma_{b_{1i}}$.

Variance components for b_{2i} and b_{3i} can be interpreted in similar fashion.

Table 1: Estimated regression coefficients (fixed effects) and standard errors for the percent body fat data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	21.3614	0.5646	37.84
time	0.4171	0.1572	2.65
$(\text{time})_+$	2.0471	0.2280	8.98

Table 2: Estimated covariance of the random effects and standard errors for the percent body fat data.

PARAMETER	ESTIMATE	SE	Z
$\text{Var}(b_{1i})$	45.9413	5.7393	8.00
$\text{Var}(b_{2i})$	1.6311	0.4331	3.77
$\text{Var}(b_{3i})$	2.7497	0.9635	2.85
$\text{Cov}(b_{1i}, b_{2i})$	2.5263	1.2185	2.07
$\text{Cov}(b_{1i}, b_{3i})$	-6.1096	1.8730	-3.26
$\text{Cov}(b_{2i}, b_{3i})$	-1.7505	0.5980	-2.93
$\text{Var}(\epsilon_i) = \sigma^2$	9.4732	0.5443	17.40

Results

Estimated intercept, $\widehat{\beta}_1 = 21.36$, has interpretation as the average percent body fat at menarche (when $t_{ij} = 0$).

Of note, actual percent body fat at menarche is not observed.

The estimate of the population mean pre-menarcheal slope, β_2, is 0.42, which is statistically significant at the 0.05 level.

This estimated slope is rather shallow and indicates that the annual rate of body fat accretion prior to menarche is less that 0.5%.

The estimate of the population mean post-menarcheal slope, $\beta_2 + \beta_3$, is 2.46 (with SE = 0.12), which is statistically significant at the 0.05 level.

This indicates that annual rate of body fat accretion is approximately 2.5%, almost six times higher than in the pre-menarcheal period.

Based on magnitude of $\widehat{\beta}_3$, relative to its standard error, slopes before and after menarche differ (at the 0.05 level).

Thus, there is evidence that body fat accretion differs before and after menarche.

Estimated variance of b_{1i} is 45.94, indicating substantial variability from girl to girl in true percent body fat at menarche, $\beta_1 + b_{1i}$.

For example, approximately 95% of girls have true percent body fat between 8.08% and 34.65% $\left(\text{i.e., } 21.36 \pm 1.96 \times \sqrt{45.94}\right)$.

Estimated variance of b_{2i} is 1.63, indicating substantial variability from girl to girl in rates of fat accretion during the pre-menarcheal period.

For example, approximately 95% of girls have changes in percent body fat before menarche between -2.09% and 2.92% $\left(\text{i.e., } 0.42 \pm 1.96 \times \sqrt{1.63}\right)$.

Estimated variance of slopes during the post-menarcheal period, $\text{Var}(b_{2i} + b_{3i})$, is 0.88 (or [1.63 + 2.75 $-2 \times$ 1.75]), indicating less variability in the slopes after menarche.

For example, approximately 95% of girls have changes in percent body fat between 0.62% and 4.30% $\left(\text{i.e., } 2.46 \pm 1.96 \times \sqrt{0.88}\right)$.

Results indicate that more than 95% of girls are expected to have increases in body fat during the post-menarcheal period.

Substantially fewer (approximately 63%) are expected to have increases in body fat during the pre-menarcheal period.

Finally, there is strong positive correlation (approximately 0.8) between annual measurements of percent body fat.

The estimated marginal correlations among annual measurements of percent body fat can be derived from the estimated variances and covariances among the random effects in Table 2.

Strength of correlation declines over time, but does not decay to zero even when measurements are taken 8 years apart (see Table 3).

Table 3: Marginal correlations (off-diagonals) among repeated measures of percent body fat between 4 years pre- and post-menarche, with estimated variances along main diagonal.

-4	-3	-2	-1	0	1	2	3	4
61.3	0.82	0.78	0.71	0.61	0.60	0.57	0.52	0.47
0.82	54.9	0.81	0.76	0.70	0.68	0.64	0.60	0.54
0.78	0.81	51.8	0.80	0.76	0.74	0.71	0.66	0.60
0.71	0.76	0.80	52.0	0.81	0.79	0.76	0.71	0.64
0.61	0.70	0.76	0.81	55.4	0.81	0.78	0.73	0.66
0.60	0.68	0.74	0.79	0.81	49.1	0.79	0.76	0.70
0.57	0.64	0.71	0.76	0.78	0.79	44.6	0.77	0.74
0.52	0.60	0.66	0.71	0.73	0.76	0.77	41.8	0.76
0.47	0.54	0.60	0.64	0.66	0.70	0.74	0.76	40.8

The mixed effects model can be used to obtain estimates of each girl's growth trajectory over time, based on the $\widehat{\beta}$'s and $\widehat{b}_i$'s.

Figure 4 displays estimated population mean growth curve and predicted (empirical BLUP) growth curves for two girls.

Note: the two girls differ in the number of measurements obtained (6 and 10 respectively).

A noticeable feature of the predicted growth curves is that there is more shrinkage towards the population mean curve when fewer data points are available.

This becomes more apparent when BLUPs are compared to ordinary least squares (OLS) estimates based only on data from each girl (see Figure 5).

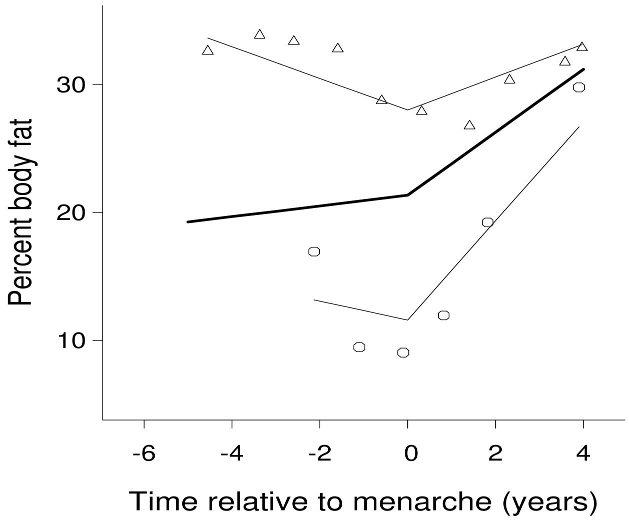

Figure 4: Population average curve and empirical BLUPs for two randomly selected girls.

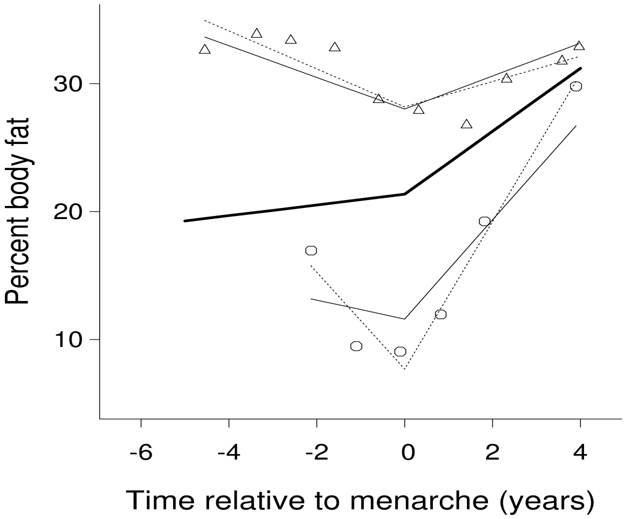

Figure 5: Population average curve, empirical BLUPs, and OLS predictions for two randomly selected girls.

Summary of Key Points

Linear mixed effects models are increasingly used for the analysis of longitudinal data.

Introduction of random effects accounts for the correlation among repeated measures and allows for heterogeneity of the variance over time, but does not change the model for $E(Y_{ij}|X_{ij})$.

The inclusion of random slopes or random trajectories induces a random effects covariance structure for $Y_{i1}, ..., Y_{in_i}$ where the variances and correlations are a function of the times of measurement.

In general, the random effects covariance structure is relatively parsimonious (e.g., random intercepts and slopes model has only four parameters, $\sigma^2_{b_1}, \sigma^2_{b_2}, \sigma_{b_1,b_2},$ and σ^2).

Linear mixed effects models are appealing because of

- their flexibility in accommodating a variety of study designs, data models and hypotheses.
- their flexibility in accommodating any degree of imbalance in the data (e.g., due to mistimed measurements and/or missing data)
- their ability to parsimoniously model the variance and correlation
- their ability to predict *individual* trajectories over time

Note 1: Tests of fixed effects rely on asymptotic normality of the fixed effects (not Y_{ij}); need reasonable (say > 30) number of subjects.

Note 2: Missing observations can be accommodated easily, validity of results depends upon assumption about missingness.

Linear Mixed Models using PROC MIXED in SAS

Table 4: Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group;
   MODEL y=group time group*time / SOLUTION CHISQ;
   RANDOM INTERCEPT time / SUBJECT=id TYPE=UN G V;
```

Table 5: Illustrative commands for obtaining the estimated BLUPs and predicted responses from model with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
PROC MIXED;
  CLASS id group;
  MODEL y=group time group*time / SOLUTION CHISQ OUTPRED=yhat;
  RANDOM INTERCEPT time / SUBJECT=id TYPE=UN SOLUTION;

PROC PRINT;
  VAR id group time y PRED;
```