Biostatistics Workshop 2008

$\sim$ Longitudinal Data Analysis $\sim$

Session 4

GARRETT FITZMAURICE

Harvard University

LINEAR MIXED EFFECTS MODELS

Motivating Example: *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.

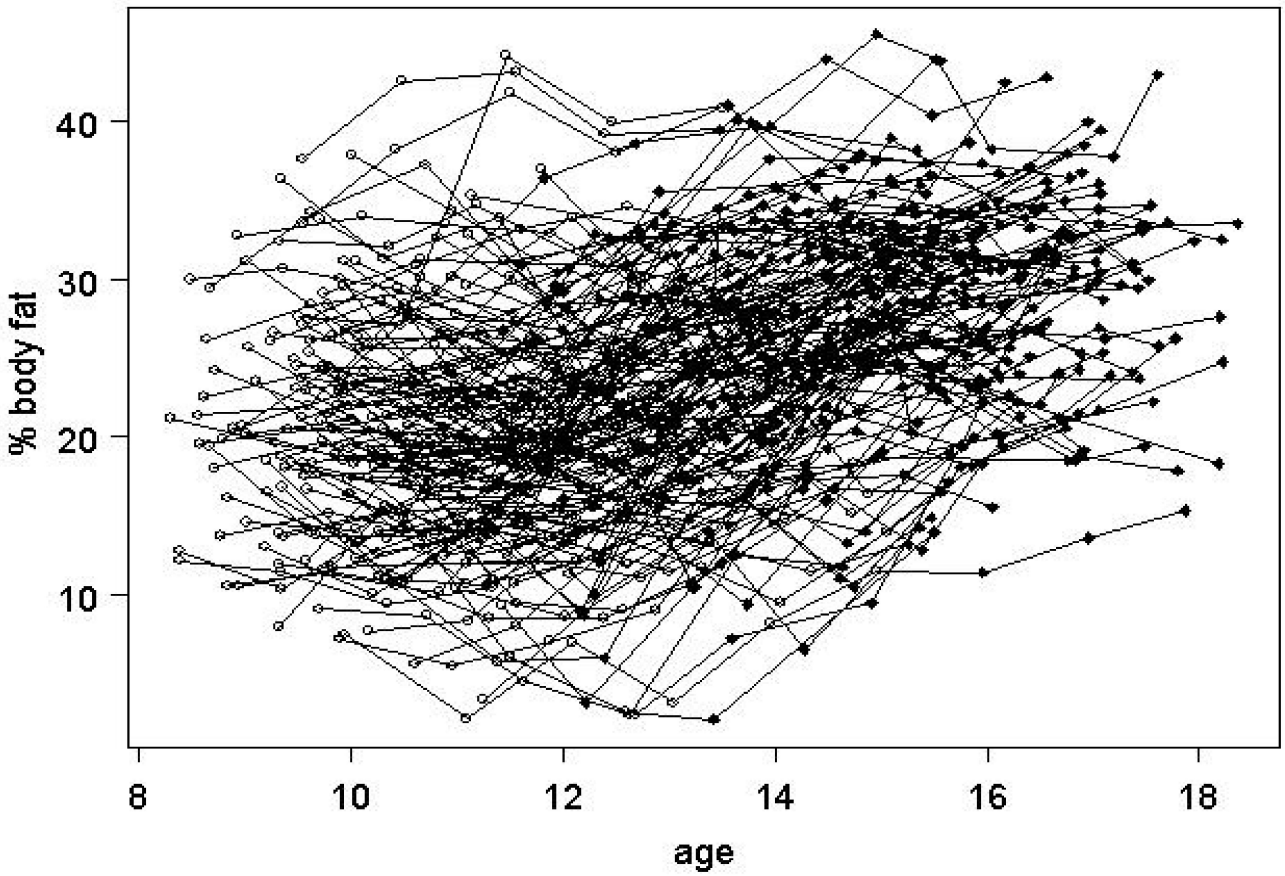

Figure 1: Timeplot of percent body fat against age (in years).

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 "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.

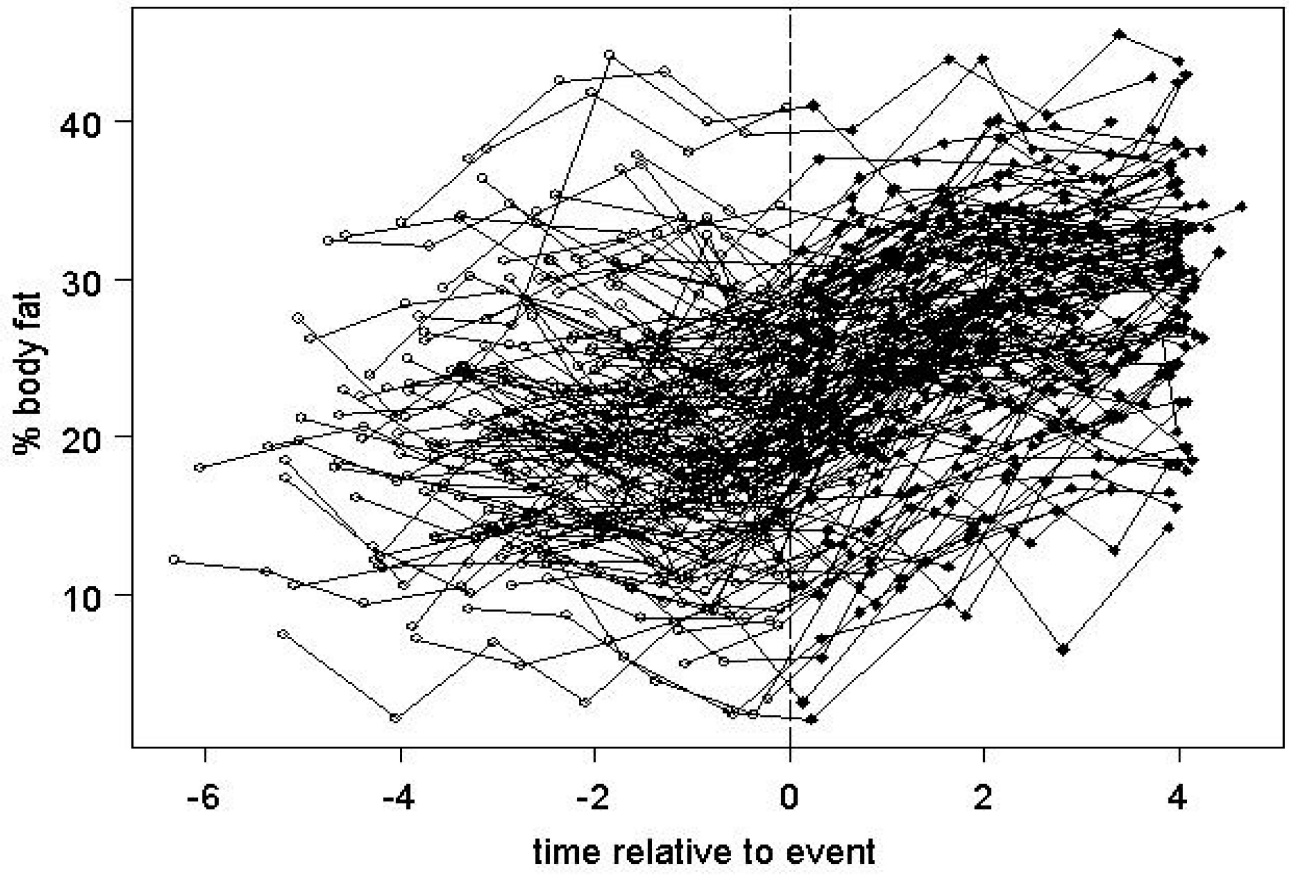

Figure 2: Timeplot of percent body fat against time, relative to age of menarche (in years).

LINEAR MIXED EFFECTS MODELS

Basic idea: Individuals in population are assumed to have their own subject-specific mean response trajectories over time.

Allow subset of the regression parameters to vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population.

Distinctive feature: mean response modelled as a combination of population characteristics (*fixed effects*) assumed to be shared by all individuals, and subject-specific effects (*random effects*) that are unique to a particular individual.

The term *mixed* denotes that model contains both fixed and random effects.

Linear Models for the Mean Response

The mean response can be modelled by a familiar regression model.

For example, with a linear trend over time, we may have

$$E(Y_{ij}|X_{ij}) = \mu_{ij} = \beta_0 + \beta_1 t_{ij}.$$

With additional covariates, this can be written more generally

$$E(Y_{ij}|X_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$$

where t_{ij}, or possibly functions of t_{ij}, have been incorporated into the covariates, e.g., $X_{ij1} = t_{ij}$, $X_{ij2} =$ treatment group indicator, and $X_{ij3} = t_{ij} \times$ treatment group indicator.

Models for Correlation: Random Intercept Model

One traditional approach for handling the correlation among repeated measures is to assume that it arises from a random subject effect.

That is, each subject is assumed to have an (unobserved) underlying level of response which persists across his/her repeated measurements.

This subject effect is treated as random and the model becomes

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij}$$

or

$$Y_{ij} = (\beta_0 + b_i) + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + \epsilon_{ij}$$

(also known as "random intercept model").

In the model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij}$$

the response for the i^{th} subject at j^{th} occasion is assumed to differ from the population mean,

$$\mu_{ij} = E(Y_{ij}|X_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$$

by a subject effect, b_i, and a within-subject measurement error, ϵ_{ij}.

Furthermore, it is assumed that

$$b_i \sim N(0, \sigma_b^2); \qquad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

and that b_i and ϵ_{ij} are mutually independent.

Figure 3 provides graphical representation of linear trend model:

$$Y_{ij} = (\beta_0 + b_i) + \beta_1 t_{ij} + \epsilon_{ij}$$

Overall mean response over time in the population changes linearly with time (denoted by the solid line).

Subject-specific mean responses for two specific individuals, subjects A and B, deviate from the population trend (denoted by the broken lines).

Individual A responds "higher" than the population average and thus has a positive b_i.

Individual B responds "lower" than the population average and has a negative b_i.

Inclusion of measurement errors, ϵ_{ij}, allows response at any occasion to vary randomly above/below subject-specific trajectories (see Figure 4).

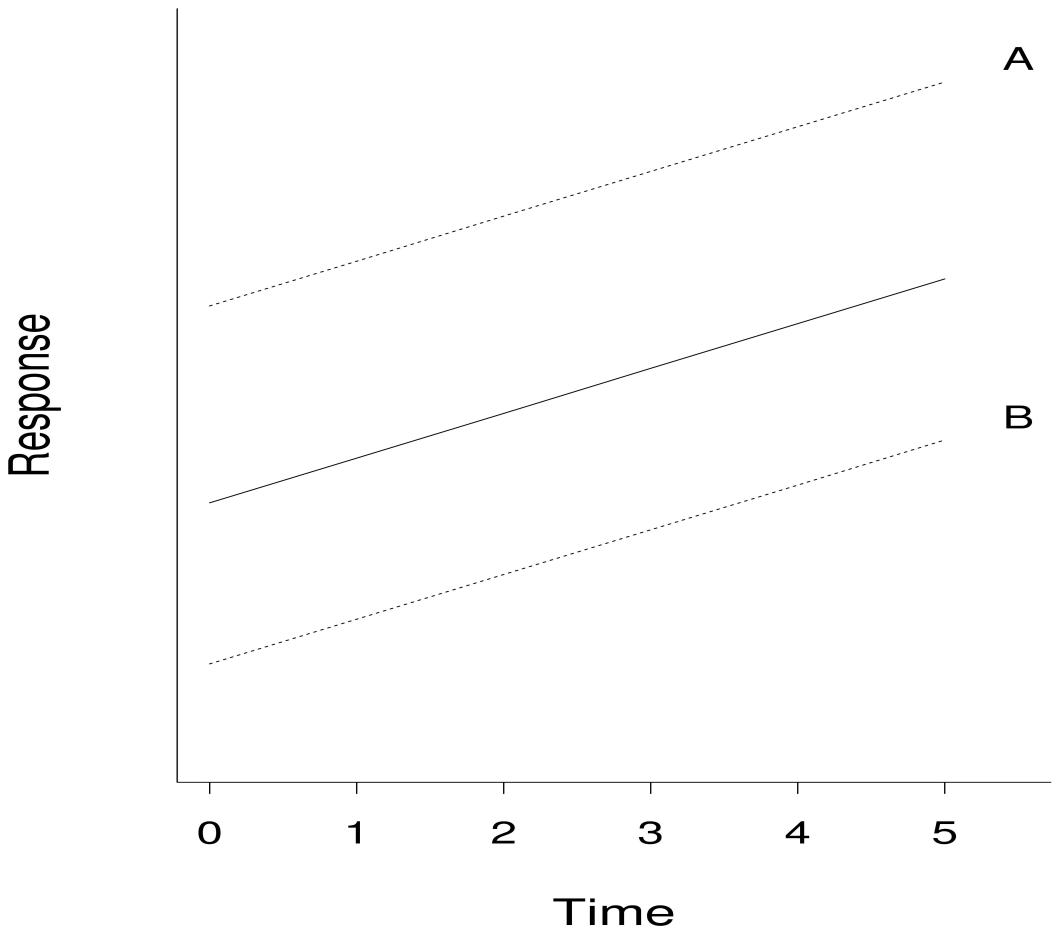

Figure 3: Graphical representation of the overall and subject-specific mean responses over time.

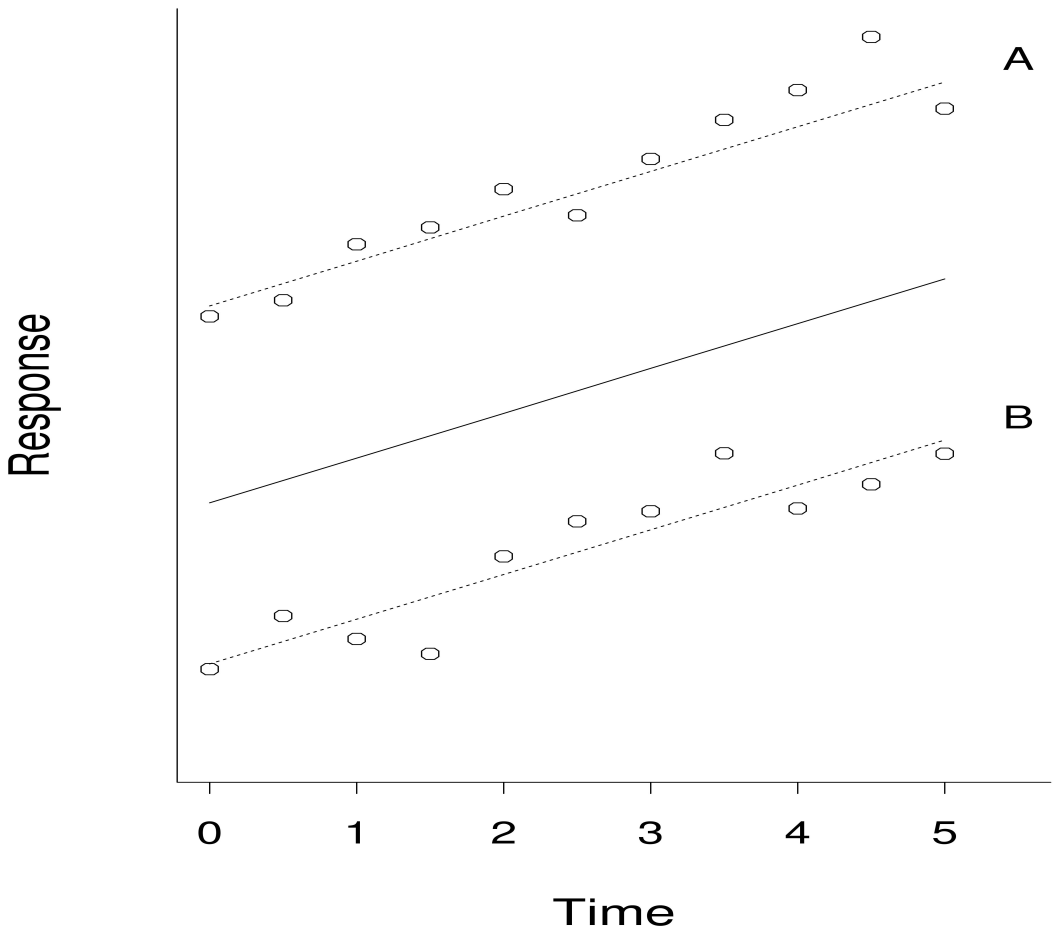

Figure 4: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Covariance/Correlation Structure

The introduction of a random subject effect induces correlation among the repeated measures.

The following "compound symmetry" covariance structure results:

$$\text{Var}(Y_{ij}) = \sigma_b^2 + \sigma_\epsilon^2$$

$$\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_b^2 \Longrightarrow \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}$$

This is the correlation among pairs of observations on the same individual.

Potential Drawback: Variances and correlations are assumed to be constant.

Solution: Allow for heterogeneity is trends over time $\Longrightarrow$ random intercepts and slopes.

Extension: Random Intercept and Slope Model

Consider a model with intercepts and slopes that vary randomly among individuals,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \quad j = 1, ..., n_i,$$

where t_{ij} denotes the timing of the j^{th} response on the i^{th} subject.

This model posits that individuals vary not only in their baseline level of response (when $t_{i1} = 0$), but also in terms of their changes in the response over time (see Figure 5).

The effects of covariates (e.g., due to treatments, exposures) can be incorporated by allowing mean of intercepts and slopes to depend on covariates.

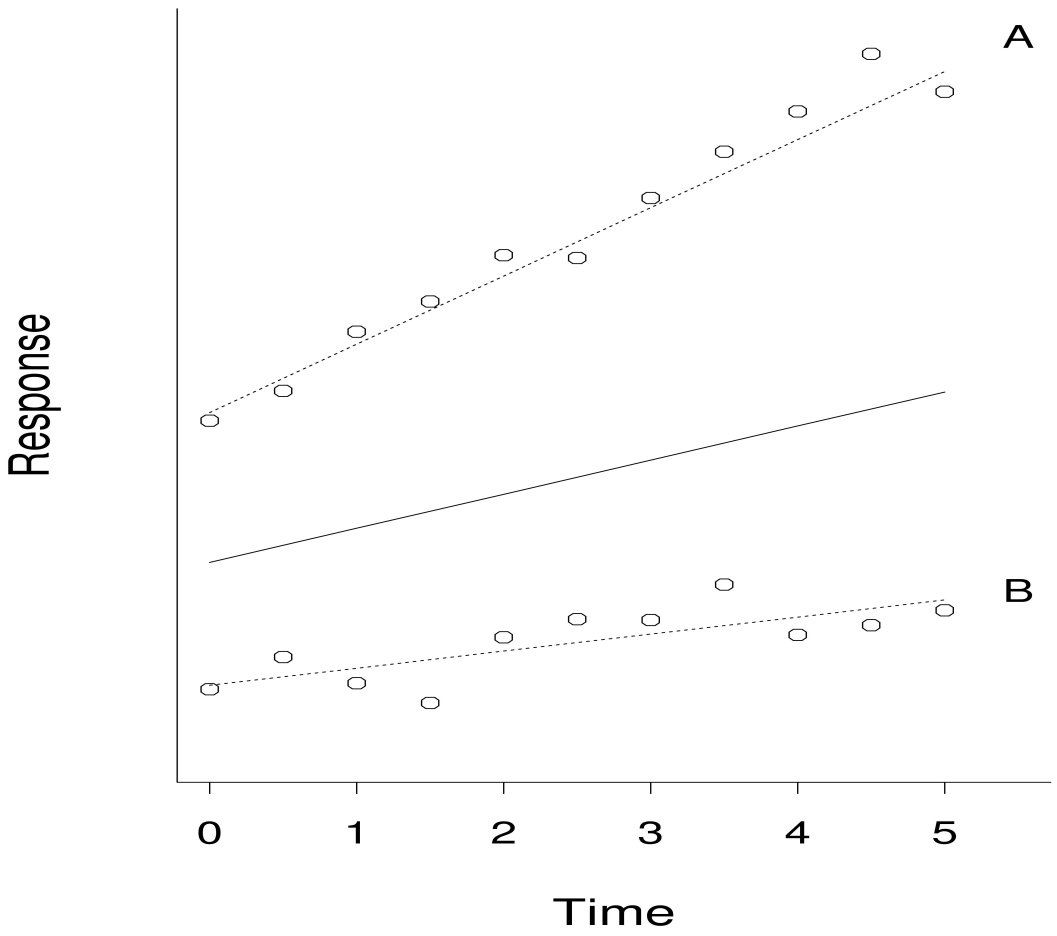

Figure 5: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

For example, consider two-group study comparing a *treatment* and a *control* group:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{trt}_i + \beta_3 t_{ij} \times \text{trt}_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij},$$

where $\text{trt}_i = 1$ if the i^{th} individual assigned to treatment group, and $\text{Group}_i = 0$ otherwise.

The model can be re-expressed as follows for the *control* group and *treatment* group respectively:

trt = 0: $Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + \epsilon_{ij},$

trt = 1: $Y_{ij} = (\beta_0 + \beta_2 + b_{0i}) + (\beta_1 + \beta_3 + b_{1i}) t_{ij} + \epsilon_{ij},$

Finally, consider the covariance induced by the introduction of random intercepts and slopes.

Assuming $b_{0i} \sim N(0, \sigma^2_{b_0})$, $b_{1i} \sim N(0, \sigma^2_{b_1})$ (with $\mathrm{Cov}(\mathrm{b}_{0i}, b_{1i}) = \sigma_{b_0,b_1}$) and $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$, then

$$\begin{aligned}
\mathrm{Var}\,(Y_{ij}) &= \mathrm{Var}\,(b_{0i} + b_{1i}t_{ij} + \epsilon_{ij}) \\
&= \mathrm{Var}(b_{0i}) + 2t_{ij}\mathrm{Cov}(b_{0i}, b_{1i}) + t^2_{ij}\mathrm{Var}(b_{1i}) + \mathrm{Var}(\epsilon_{ij}) \\
&= \sigma^2_{b_0} + 2t_{ij}\sigma_{b_0,b_1} + t^2_{ij}\sigma^2_{b_1} + \sigma^2_\epsilon.
\end{aligned}$$

Similarly, it can be shown that

$$\mathrm{Cov}\,(Y_{ij}, Y_{ik}) = \sigma^2_{b_0} + (t_{ij} + t_{ik})\,\sigma_{b_0,b_1} + t_{ij}t_{ik}\sigma^2_{b_1}.$$

Thus, in this mixed effects model for longitudinal data the variances and correlations (covariance) are expressed as an explicit function of time, t_{ij}.

Linear Mixed Effects Model

Can allow any subset of the regression parameters to vary randomly.

Using vector notation, the linear mixed effects model can be expressed as

$$Y_{ij} = X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij},$$

where b_i is a $(q \times 1)$ vector of random effects and Z_{ij} is the vector of covariates linking the random effects to Y_{ij}.

Note: Components of Z_{ij} are a subset of the covariate in X_{ij}, e.g., in random intercepts and slopes model $X_{ij} = [1 \;\; t_{ij} \;\; \text{trt}_i \;\; t_{ij} * \text{trt}_{ij}]$ and $Z_{ij} = [1 \;\; t_{ij}]$.

Specifically, any component of β can be allowed to vary randomly by simply including corresponding covariate in Z_{ij}.

The random effects, b_i, are assumed to have a multivariate normal distribution with mean zero and covariance matrix denoted by D.

Estimation: Maximum Likelihood

ML estimator of $\beta_0, \beta_1, ..., \beta_p$ is the *generalized least squares* (GLS) estimator and depends on covariance among the repeated measures.

This is a generalization of the ordinary least squares (OLS) estimator used in standard linear regression.

In general, there is no simple expression for ML estimator of the covariance components - requires iterative techniques.

Because ML estimation of covariance is known to be biased in small samples, use *restricted* ML (REML) estimation instead.

Prediction of Random Effects

In many applications, inference is focused on fixed effects, $\beta_0, \beta_1, ..., \beta_p$.

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

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

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

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij},$$

where $\text{Var}\,(b_i) = \sigma_b^2$ and $\text{Var}\,(\epsilon_{ij}) = \sigma_\epsilon^2$.

It can be shown that the BLUP for b_i is:

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

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_ϵ^2.

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 "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.

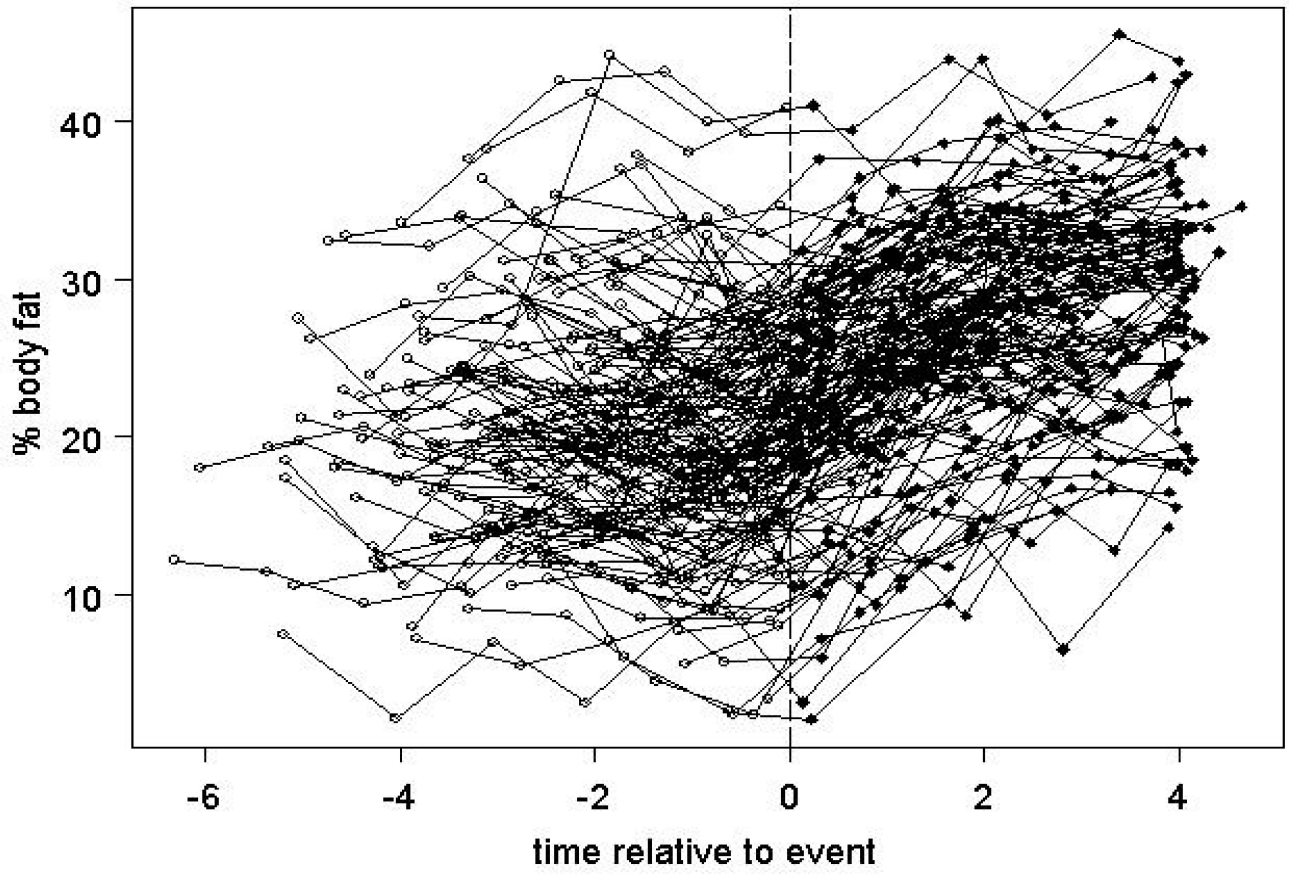

Figure 6: Timeplot of percent body fat against time, relative to age of menarche (in years).

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 7).

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 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).

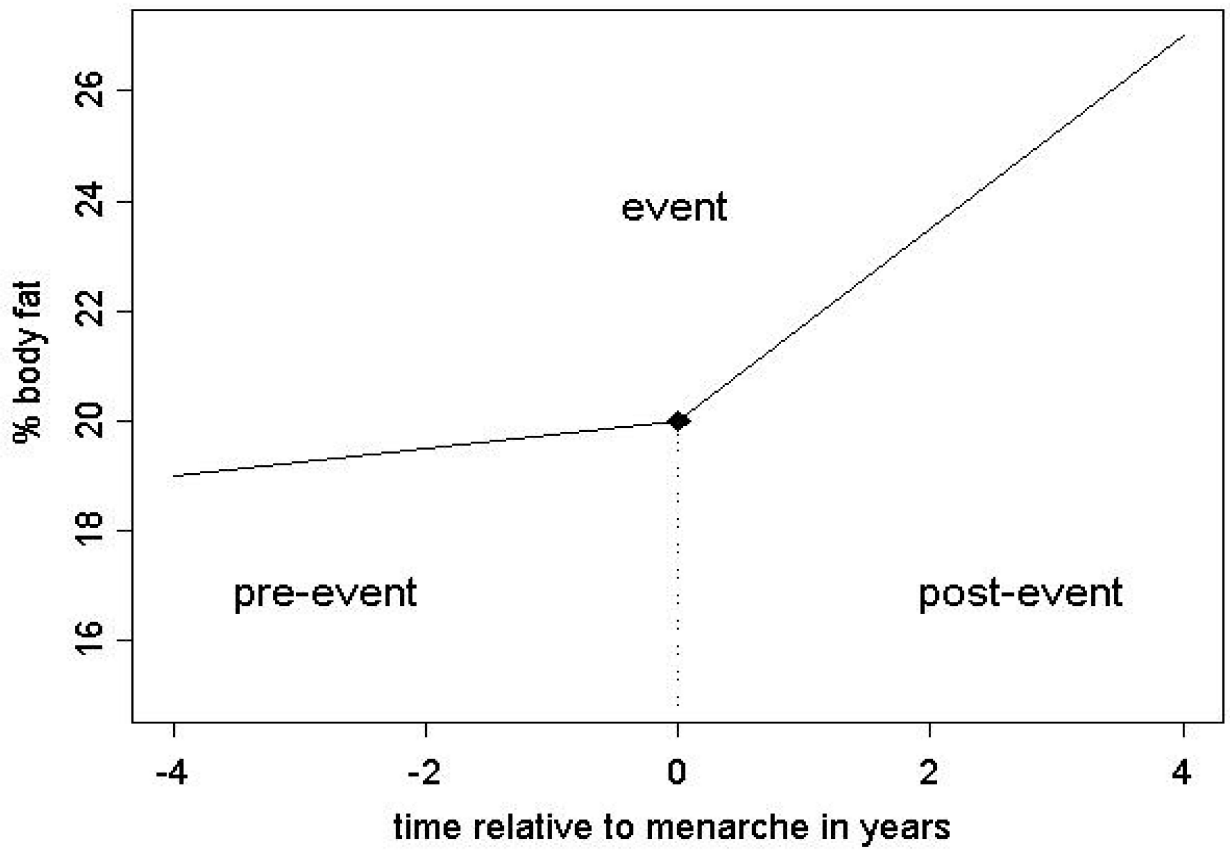

Figure 7: Graphical representation of piecewise linear trajectory.

We consider the following linear mixed effects model

$$E(Y_{ij}|b_i) = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij})_+ + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij})_+,$$

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_0 is the average %body fat at menarche (when $t_{ij} = 0$).

The slope β_1 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 $(\beta_1 + \beta_2)$.

Goal: Assess whether population slopes differ before and after menarche, i.e., $H_0 : \beta_2 = 0$.

Similarly, $(\beta_0 + b_{0i})$ is intercept for i^{th} subject and is the true %body fat at menarche (when $t_{ij} = 0$).

$(\beta_1 + b_{1i})$ 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 $[(\beta_1 + \beta_2) + (b_{1i} + b_{2i})]$.

Interpretation of variance components:

Recall that the subject-specific intercepts, $(\beta_0 + b_{0i})$, have mean β_0 and variance $\sigma^2_{b_{0i}}$.

Furthermore, since $b_{0i} \sim N(0, \sigma^2_{b_{0i}})$ this implies that $(\beta_0 + b_{0i}) \sim N(\beta_0, \sigma^2_{b_{0i}})$.

Under the assumption of normality, we expect 95% of the subject-specific intercepts, $(\beta_0 + b_{0i})$, to lie between: $\beta_0 \pm 1.96 \times \sigma_{b_{0i}}$.

Variance components for b_{1i} and b_{2i} 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
$\mathrm{Var}(b_{0i})$	45.9413	5.7393	8.00
$\mathrm{Var}(b_{1i})$	1.6311	0.4331	3.77
$\mathrm{Var}(b_{2i})$	2.7497	0.9635	2.85
$\mathrm{Cov}(b_{0i}, b_{1i})$	2.5263	1.2185	2.07
$\mathrm{Cov}(b_{0i}, b_{2i})$	-6.1096	1.8730	-3.26
$\mathrm{Cov}(b_{1i}, b_{2i})$	-1.7505	0.5980	-2.93
$\mathrm{Var}(e_i) = \sigma_e^2$	9.4732	0.5443	17.40

Results

Estimated intercept, $\widehat{\beta}_0 = 21.36$, has interpretation as the average percent body fat at merarche (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_1, 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 is less that 0.5%.

The estimate of the population mean post-menarcheal slope, $\beta_1 + \beta_2$, 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 $\widehat{\beta}_2$, 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_{0i} is 45.94, indicating substantial variability from girl to girl in true percent body fat at menarche, $\beta_0 + b_{0i}$.

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_{1i} is 1.6, 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 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, $\text{Var}(b_{1i} + b_{2i})$, 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 8 displays estimated population mean growth curve and predicted (empirical BLUP) growth curves for two girls.

Note: 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 9).

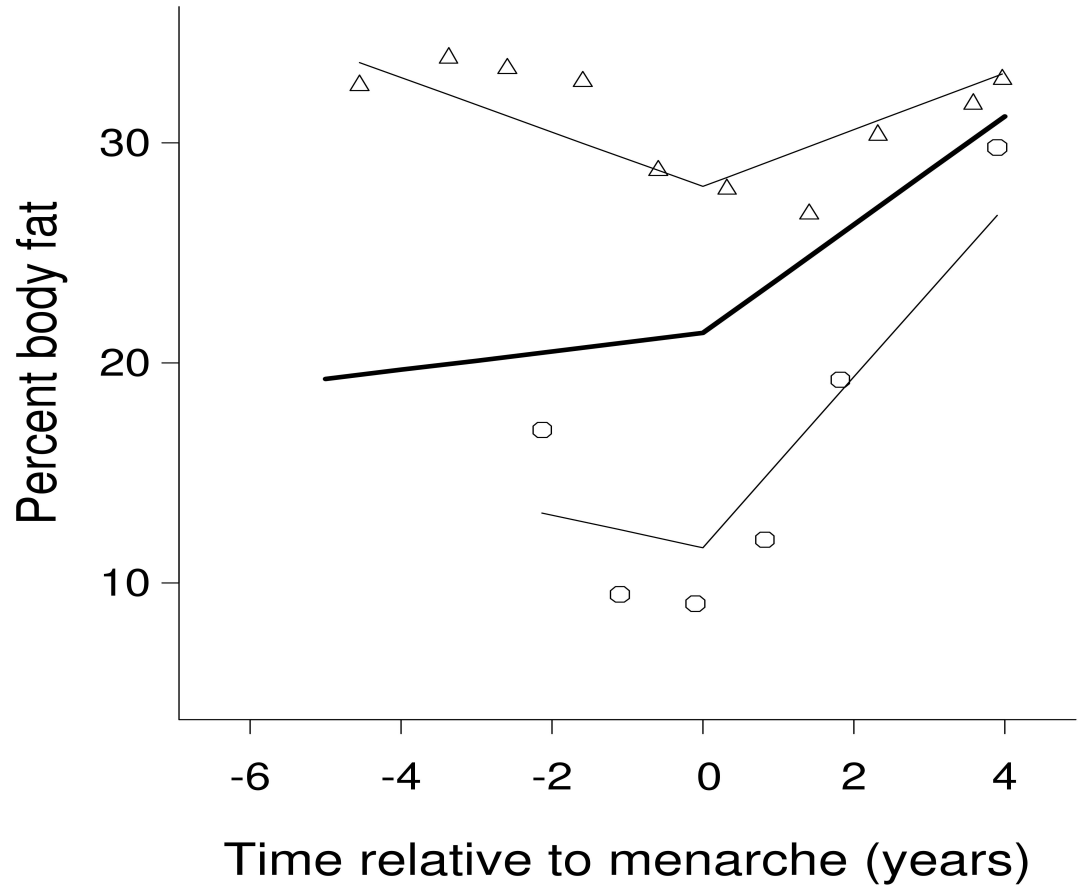

Figure 8: Population average curve and empirical BLUPs for two randomly selected girls.

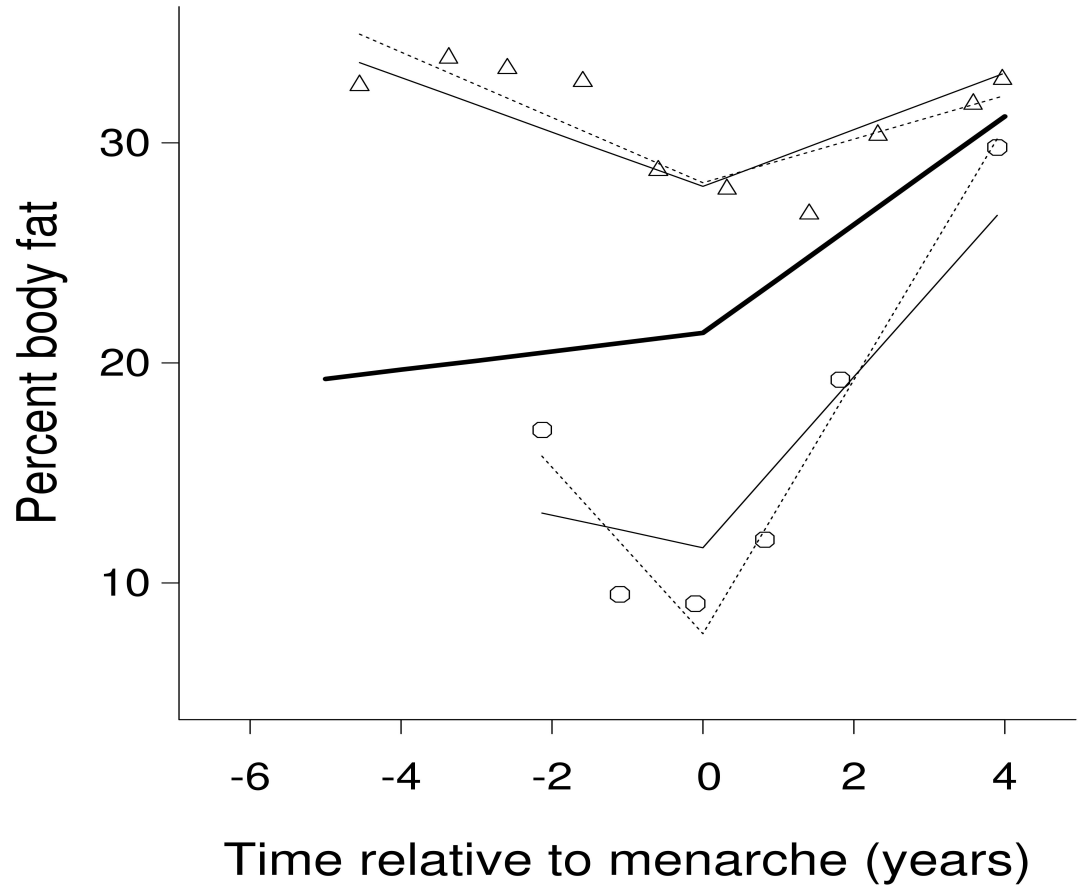

Figure 9: Population average curve, empirical BLUPs, and OLS predictions for two randomly selected girls.

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, $\sigma^2_{b_0}, \sigma^2_{b_1}, \sigma_{b_0,b_1}$, and σ^2_e).

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

Note 2: Missing observations can be accommodated easily, validity of results depends upon assumption about missingness (see Session 7).

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;
```

Linear Mixed Models using xtmixed in Stata

Table 6: Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, using xtmixed in Stata. The predict postestimation command is used for obtaining the estimated BLUPs.

```
. generate grp_time=group*time

. xtmixed y group time grp_time || id: time, cov(unstr) variance

. predict b1 b0, reffects
```

FURTHER READING

Diggle, P.J., Heagerty, P., Liang, K-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press. (See Chapter 5).

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*. Wiley. (See Chapter 8).

Naumova, E.N., Must, A. and Laird, N.M. (2001). Evaluating the impact of "critical periods" in longitudinal studies of growth using piecewise mixed effects models. *International Journal of Epidemiology*, **30**, 1332–1341.

Cnaan, A., Slasor, P. and Laird, N.M. (1997). Tutorial in Biostatistics: Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, **16**, 2349–2380.