Random Coefficient Model

(another way to consider the repeated measures study)

Population-level regression line with individual-level regression lines

The focus is on individual trajectories, but the mean response is still of interest.

This type of modeling can address research questions related to an individual's trajectory, where our previous models may not be set-up for that type of question.
Consider the model

\[ y_{ij} = b_0 + b_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + e_{ij} \]

where

\[ b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D_{2 \times 2}) \]

\[ e_{ij} \sim N(0, \sigma^2) \]

An individual is modeled with its own line:

\[ E[ y_{ij} | b_{0i}, b_{1i}] = (b_0 + b_{0i}) + (b_1 + b_{1i}) x_{ij} \]

There is also an overall, average trajectory:

\[ E[ y_{ij}] = b_0 + b_1 x_{ij} \]

Within-subject variation is modeled as fluctuations about the subject-specific line using \( e_{ij} \).

Between-subject variation is modeled through \( D_{2 \times 2} \). How does each individual trajectory vary from the population mean trajectory? How different are the individual slopes & intercepts? How do the slope & intercept covary? (e.g. do large slopes & intercepts tend to occur together?)
Letting $Z$ represent the random effects design matrix, 
\[ Y = X\beta + Z\gamma + \epsilon \]
random coefficients for each individual ($b_0, b_1$) are in this vector. 
Let $Y_i = b_0; = (b_0, b_1)^T$

For the $j$th observation on the $i$th subject 
we have the "design vector" of $Z_{ij} = (1, X_{ij})$
and the "individual design matrix"
\[
Z_i = \begin{pmatrix}
1 & X_{i1} \\
1 & X_{i2} \\
1 & X_{i3} \\
1 & X_{in}
\end{pmatrix}
\]

This is how we can estimate subject-specific regression lines.

In the present modeling, we have 
\[ \text{Var}(Y_i) = \text{Var}(\gamma_i) = D = \begin{pmatrix} d_{11} & d_{12} \\
 d_{21} & d_{22} \end{pmatrix} \]
and \[ \text{Var}(\epsilon_i) = \text{Var}(\gamma_i) = \text{Var}(\epsilon_i) = \sigma^2 I_n; \]

Again, \[ \text{Var}(\gamma_i) = \Sigma \]
is a block diagonal.

At the individual level, 
\[ \text{Var}(Y_i) = Z_iDZ_i^T + R_i \]
\[ \text{Var}(\gamma_i | b_i) = \sigma^2 I_n; \]
Correlation is induced through the random effects random effects are no longer independent.
Slope & intercept are allowed to be correlated.
The same correlation is used for all individuals.
Example: Let $x$ be time for $t = t_1, t_2, t_3$.

\[ Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix}, \quad x_i = \begin{pmatrix} 1 \\ t_1 \\ t_2 \\ t_3 \end{pmatrix}, \quad z_i = \begin{pmatrix} 1 \\ t_1 \\ t_2 \\ t_3 \end{pmatrix} \]

\[ \text{Var}(Y_{ij}) = \text{Var}(z_i' x_i + e_i) = z_i' D z_i + \sigma^2 I_{n_i} \]

\[ \text{Var}(Y_{ij}) = d_{11} + t_j^2 d_{22} + 2t_j d_{21} + \sigma^2 \]

The diagonal element for subject $i$ at observation $j$.

thus, variance can change over time, function of $t_j$.

\[ \text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(b_{0i} + t_j b_{1i} + e_{ij}, b_{0i} + t_k b_{1i} + e_{ik}) \]

\[ = \text{Cov}(b_{0i}, b_{0i}) + \text{Cov}(b_{0i}, t_k b_{1i}) + \text{Cov}(t_j b_{1i}, b_{0i}) + \text{Cov}(t_j b_{1i}, t_k b_{1i}) \]

\[ = d_{11} + t_k d_{21} + t_j d_{21} + t_j t_k d_{22} \]

\[ = d_{11} + t_j t_k d_{22} + (t_j + t_k) d_{21} \]

\[ = (1, t_j) \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} \begin{pmatrix} 1 \\ t_k \end{pmatrix} \]

\[ = z_i' \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} z_i' \]
As a comment, note that this model could also be stated as:

\[ Y_{ij} = b_{0i}^* + b_{1i}^* X_{ij} + e_{ij} \]

with \( \begin{pmatrix} \hat{b}_{0i}^* \\ \hat{b}_{1i}^* \end{pmatrix} \sim N \left( \begin{pmatrix} B_0 \\ B_1 \end{pmatrix}, D \right) \)

The covariance matrix depends on the specific times of observation (when \( X \) is a time, that is) through the matrix \( \mathbf{Z}_i \).

Thus, if different subjects are observed at different times, this information is automatically incorporated into the model.

References:

* Random-effects Models for Longitudinal Data
* Notes from Marie Davidian, NC State, Applied Longitudinal Data Analysis
Random Coefficient Model (a.k.a. multilevel model)

(Adapted from UCLA Statistical Computing Seminars)

School math achievement scores

The data file consists of 7185 students nested in 160 schools. The outcome variable of interest is student-level math achievement score (MATHACH). Variable SES is social-economic-status of a student and therefore is a student-level variable. Variable MEANSES is the group mean of SES and therefore is a school-level variable. Both SES and MEANSES are centered at the grand mean (they both have means of 0). Variable SECTOR is an indicator variable indicating if a school is public or catholic and is therefore a school-level variable. There are 90 public schools (SECTOR=0) and 70 catholic schools (SECTOR=1) in the sample.

A segment of the data file:

<table>
<thead>
<tr>
<th>SCHOOL</th>
<th>MATHACH</th>
<th>SES</th>
<th>MEANSES</th>
<th>SECTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1296</td>
<td>6.588</td>
<td>-0.178</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>11.026</td>
<td>0.392</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>7.095</td>
<td>-0.358</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>12.721</td>
<td>-0.628</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>5.520</td>
<td>-0.038</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>7.353</td>
<td>0.972</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>7.095</td>
<td>0.252</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>9.999</td>
<td>0.332</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1296</td>
<td>10.715</td>
<td>-0.308</td>
<td>-0.420</td>
<td>0</td>
</tr>
<tr>
<td>1308</td>
<td>13.233</td>
<td>0.422</td>
<td>0.534</td>
<td>1</td>
</tr>
<tr>
<td>1308</td>
<td>13.952</td>
<td>0.562</td>
<td>0.534</td>
<td>1</td>
</tr>
<tr>
<td>1308</td>
<td>13.757</td>
<td>-0.058</td>
<td>0.534</td>
<td>1</td>
</tr>
<tr>
<td>1308</td>
<td>13.970</td>
<td>0.952</td>
<td>0.534</td>
<td>1</td>
</tr>
</tbody>
</table>
1. Is there a significant school effect? (Model that disregards covariates).

Let \( i = \text{student} \) and \( j = \text{school} \).

\[
y_{ij} = \mu + u_{0j} + r_{ij} \quad \text{with} \quad u_{0j} \overset{iid}{\sim} N(0, \sigma_r^2) \quad \text{and} \quad r_{ij} \overset{iid}{\sim} N(0, \sigma^2)
\]

where \( u_{0j} \) and \( r_{ij} \) independent.

This looks like our usual 1-way ANOVA random effects model, or some may call it the *unconditional means model*.

As a random coefficient model with a random intercept:

```plaintext
proc mixed data = rc.hsb12 covtest noclprint;
    class school;
    model mathach = / solution;
    random intercept / subject = school;
run;
```

**Covariance Parameter Estimates**

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Z Value</th>
<th>Pr &gt; Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>SCHOOL</td>
<td>8.6097</td>
<td>1.0778</td>
<td>7.99</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>39.1487</td>
<td>0.6607</td>
<td>59.26</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**Unconditional Means Model**

**The Mixed Procedure**

**Solution for Fixed Effects**

| Effect       | Estimate | Standard Error | DF | t Value | Pr > |t| |
|--------------|----------|----------------|----|---------|------|---|
| Intercept    | 12.6370  | 0.2443         | 159| 51.72   | <.0001 |

As a mixed model with a random SCHOOL effect:

```plaintext
proc mixed data=rc.hsb12 covtest;
    class school;
    model mathach=/ddfm=satterth solution;
    random school;
run;
```
Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Z</th>
<th>Pr &gt; Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCHOOL</td>
<td>8.6097</td>
<td>1.0778</td>
<td>7.99</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>39.1487</td>
<td>0.6607</td>
<td>59.26</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect   | Estimate | Standard Error | DF  | t Value | Pr > |t| |
|----------|----------|----------------|-----|---------|------|-----|
| Intercept| 12.6370  | 0.2443         | 157 | 51.72   | <.0001|

2. Do schools with high socio-economic status (SES) tend to have high math achievement scores (MATHACH)?

Model 2: Include fixed effect for MEANSES (continuous variable) while accounting for the random SCHOOL effect (as compound symmetry).

Let \(i=\text{student and } j=\text{school.}\)

\[ y_{ij} = \mu + \beta_1 x_j + u_{0j} + r_{ij} \quad \text{with} \quad u_{0j} \sim iid N(0, \sigma_2^2) \quad \text{and} \quad r_{ij} \sim iid N(0, \sigma^2) \]

where \(u_{0j}\) and \(r_{ij}\) independent.

This model assumes a linear relationship with MEANSES from the schools, allows for a separate random intercept for each school, and assumes constant variance across schools. The correlation structure is compound symmetry.

```r
proc mixed data = rc.hsb12 covtest noclprint;
   class school;
   model mathach = meanse / solution ddfm = satterth;
   random intercept / subject = school;
run;
```

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Z</th>
<th>Pr &gt; Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>SCHOOL</td>
<td>2.6357</td>
<td>0.4036</td>
<td>6.53</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>39.1578</td>
<td>0.6608</td>
<td>59.26</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
## Solution for Fixed Effects

| Effect    | Estimate | Standard Error | DF | t Value | Pr > |t| |
|-----------|----------|----------------|----|---------|------|---|
| Intercept | 12.6495  | 0.1492         | 154| 84.77   | <.0001 |
| MEANSES   | 5.8635   | 0.3613         | 154| 16.23   | <.0001 |

If we compare the estimated SCHOOL variance component (representing variation between schools) from model 1 and model 2, the magnitude decreases greatly (from 8.6097 to 2.6357). This means that the variable meanSES explains a large portion of the school-to-school variation in mean math achievement.

Do school achievement means still vary significantly once MEANSES is controlled? From the output of Covariance Parameter Estimates, we see that the test that between variance is zero is highly significant. Therefore, we conclude that after controlling for MEANSES, significant variation among school mean math achievement still remains to be explained.

Again, the same results would be found from the following SAS code:

```sas
proc mixed data = rc.hsb12 covtest noclprint;
  class school;
  model mathach = meanses / solution ddfm=satterth;
  random school;
run;
```

3. What if we include the student-level socio-economic level (SES)? (Random coefficient model - random intercept and random slope).

\[ y_{ij} = \mu + \beta_1 x_{ij}^* + u_{0j} + u_{1j} x_{ij}^* + r_{ij} \]

with

\[
\begin{pmatrix}
  u_{0j} \\
  u_{1j}
\end{pmatrix} \sim iid N(0, D) \quad \text{and} \quad r_{ij} \sim iid N(0, \sigma^2)
\]

where \((u_{0j}, u_{1j})\) and \(r_{ij}\) independent.

\(x^*\) is the centered covariate wrt school MEANSES.

This model allows for a separate regression line for each school (both intercept and slope). It assumes constant variance for each school, but this model can allow for heteroscedasticity at the marginal level (because it allows for separate lines by school which can ‘spread out’ as ses increases, for instance).
The motivating questions for this model:

(a) What would be the average of the 160 regression equations (both intercept and slope)?
(b) How much do the regression equations vary from school to school?
(c) What is the correlation between the intercepts and slopes?

```
data rc.hsb12; set rc.hsb12;
   cses = ses - meanses;
run;

proc mixed data = rc.hsb12 noclprint covtest noitprint;
   class school;
   model mathach = cses / solution ddfm = satterth notest;
   random intercept cses / subject = school type = un gcorr;
run;
```

**Covariance Parameter Estimates**

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Z Value</th>
<th>Pr Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>SCHOOL</td>
<td>8.6769</td>
<td>1.0786</td>
<td>8.04</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>SCHOOL</td>
<td>0.05075</td>
<td>0.4062</td>
<td>0.12</td>
<td>0.9006</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>SCHOOL</td>
<td>0.6940</td>
<td>0.2808</td>
<td>2.47</td>
<td>0.0067</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>36.7006</td>
<td>0.6258</td>
<td>58.65</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**Estimated G Correlation Matrix**

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>SCHOOL</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1224</td>
<td>1.0000</td>
<td>0.02068</td>
</tr>
<tr>
<td>2</td>
<td>cses</td>
<td>1224</td>
<td>0.02068</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Solution for Fixed Effects**

| Effect  | Estimate | Standard Error | DF  | t Value | Pr > |t| |
|---------|----------|----------------|-----|---------|-------|-----|
| Intercept | 12.6493  | 0.2445         | 157 | 51.75   | <.0001 |
| cses     | 2.1932   | 0.1283         | 155 | 17.10   | <.0001 |
In the output of Covariance Parameter Estimates, the parameter corresponding to UN(2,2) is the variability in slopes of cses. The estimate is 0.6940 with standard error 0.2808. That yields a p-value of 0.0067 for 1-tailed test. The test being significant tells us that we can not accept the hypothesis that there is no difference in slopes among schools.

NOTE: When you have a longitudinal study and the times are not uniform, the random coefficients model is one way to model the data without considering observations to be ‘missing’.