Faculty of Health Sciences

Variance components and LMMs
Analysis of repeated measurements, 4th December 2014

Julie Lyng Forman & Lene Theil Skovgaard
Department of Biostatistics, University of Copenhagen

Topics for today

Leftover from 28/11:
▶ Rest of random regression example.

New concepts for today:
▶ random effects
▶ multi-level models
▶ linear mixed models (LMMs)
for clustered data and repeated measurements in general

Suggested reading:
▶ Fitzmaurice et al. (2011): chapters 8, 21, 22.
▶ Paper by Bland and Altman
(see section on comparison of measurement methods)

Outline

Random regression

General repeated measurements
Random effects ANOVA (the two-level model)
Variance components in general
Zero estimates of variance components
Linear mixed models (LMMs)
Comparing measurement methods

Example: Calcium supplements
Bone mineral density (\(\frac{g}{cm^2}\)) in 112 11-year old girls, randomized to either calcium or placebo.

Follow-up: every 6 months, 5 visits in total including baseline
Random regression

We let each girl have her own level $A_i$ and her own slope $B_i$.

We assume these individual 'parameters', $A_i$ and $B_i$,
- So-called random effects
follow a bivariate normal distribution in the population

$$\begin{pmatrix} A_i \\ B_i \end{pmatrix} \sim N_2 \left( \begin{pmatrix} \alpha_{g(i)} \\ \beta_{g(i)} \end{pmatrix}, \begin{pmatrix} \tau_a^2 & \omega_{ab} \\ \omega_{ab} & \tau_b^2 \end{pmatrix} \right)$$

The covariance is the so-called G-matrix:
- it describes the population variance of the lines, i.e. the inter-individual variation.

PROC MIXED: random regression

PROC MIXED DATA=calcium;
CLASS grp girl;
MODEL bmd=grp visit grp*visit / SOLUTION DDFM=SATTERTHWAITE;
RANDOM intercept visit / TYPE=UN SUBJECT=girl(grp) G;
RUN;

Individual intercepts and slopes must be specified in the random-statement.
- Here visit is used as a continuous covariate, and the intercept refers to visit=0.
- Note that type=un refers to a unstructured specification of the G-matrix. If it is omitted, we may experience convergence problems and sometimes totally incomprehensible results.
- Option g asks that the estimated G-matrix be printed.

Output from random regression

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>grp</th>
<th>girl</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>C</td>
<td>101</td>
<td>0.004105</td>
<td>3.733E-6</td>
</tr>
<tr>
<td>2</td>
<td>visit</td>
<td>C</td>
<td>101</td>
<td>3.733E-6</td>
<td>0.000048</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect | grp | Estimate | Error Estimate | DF | t Value | Pr > |t| |
|--------|-----|----------|----------------|----|---------|------|-----|
| Intercept | C   | 0.8471   | 0.008665       | 110 | 97.98   | <.0001|
| grp     | C   | 0.007058 | 0.01234        | 110 | 0.57    | 0.5685|
| grp     | F   | 0        | 0              |    |         |      |    |
| visit   | C   | 0.02242  | 0.001098       | 96.8| 20.42   | <.0001|
| visit*grp | C  | 0.004494 | 0.001571       | 96.4| 2.86    | 0.0052|
| visit*grp | F  | 0        | 0              |    |         |      |    |

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>grp</td>
<td>1</td>
<td>110</td>
<td>985.55</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>visit</td>
<td>1</td>
<td>96.4</td>
<td>985.55</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>visit*grp</td>
<td>1</td>
<td>96.4</td>
<td>8.18</td>
<td>0.0052</td>
</tr>
</tbody>
</table>

We find an extra increase in BMD of $0.0045 \ (0.0016) \ \text{g/cm}^3$
per half year, when giving calcium supplement.

Nonequidistant time points

- The girls are only seen approximately twice a year.
- Perhaps we get better estimates of the slopes when replacing visit with the actual age of the girl.
Random regression, using actual age

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>grp</th>
<th>girl</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>C</td>
<td>101</td>
<td>0.004215</td>
<td>0.000095</td>
</tr>
<tr>
<td>2</td>
<td>age11</td>
<td>C</td>
<td>101</td>
<td>0.000095</td>
<td>0.000180</td>
</tr>
</tbody>
</table>

**Solution for Fixed Effects**

| Effect   | grp | Estimate | Error | DF  | t Value | Pr > |t| |
|----------|-----|----------|-------|-----|---------|-------|
| Intercept| C   | 0.8667   | 0.0087| 110 | 99.75   | <.0001|
| grp      | P   | 0.0111   | 0.0124| 110 | 0.90    | 0.3715|
| age11    | C   | 0.0453   | 0.0022| 96  | 21.05   | <.0001|
| age11*grp| P   | 0.0089   | 0.0031|      |         |       |

In this model, we quantify the effect of a calcium supplement to $0.0089 (0.0031) \text{ g/cm}^3$ per year.

**Results from random regression**

<table>
<thead>
<tr>
<th>Group</th>
<th>Level at age 11</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.8667 (0.0087)</td>
<td>0.0453 (0.0022)</td>
</tr>
<tr>
<td>C</td>
<td>0.8778 (0.0088)</td>
<td>0.0542 (0.0022)</td>
</tr>
</tbody>
</table>

Difference: 0.0111 (0.0124) 0.0089 (0.0031)

P 0.37 0.0048

I.e. **steeper slopes** than when visit was used as the time-variable.

- Due to **quantification** (per year vs per 1/2 year)
- Due to **bias reduction** (visit is a proxy for age, and measurement error in the independent variable causes bias towards the null)

**Modeling the covariance**

Random regression implies a particular covariance pattern.

- Does this fit the data well?

**No benchmark for model comparisons:**

- An **unstructured covariance** cannot be estimated from non-equidistant data!

Instead, non-nested models can be compared using Akaike's information criterion (AIC) which balances goodness of fit against model complexity.

- **Smaller values of AIC** indicates a **better model fit**.

**Non-equidistant covariance patterns**

In case subject are measured at individual or otherwise **non-equally spaced** time points only a limited number of stationary covariance pattern models are available:

- The variance is **constant over time**.
- The correlation depend only on the time-distance between the observations.

```
proc mixed
  type= CS
  Cov(Y_{ij}, Y_{ik})
  no. param

<table>
<thead>
<tr>
<th>proc mixed</th>
<th>Cov(Y_{ij}, Y_{ik})</th>
<th>no. param</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>\sigma^2[I{j = k} + \rho \cdot I{j \neq k}]</td>
<td>2</td>
</tr>
<tr>
<td>SP (POW)(ctime)</td>
<td>\sigma^2 \rho \cdot t_{ij} &amp; - t_{ik}</td>
<td>2</td>
</tr>
<tr>
<td>SP (GAU)(ctime)</td>
<td>\sigma^2 e^{-</td>
<td>t_{ij} - t_{ik}</td>
</tr>
<tr>
<td>SP (LIN)(ctime)</td>
<td>\sigma^2 (1 - \rho</td>
<td>t_{ik} - t_{ij}</td>
</tr>
</tbody>
</table>
```

The ctime-variable must be a **numerical variable** in SAS.
Tests of treatment effect

Comparison of estimates for different covariance structures:

<table>
<thead>
<tr>
<th>Covariance structure</th>
<th>AIC</th>
<th>Cov.par.</th>
<th>Difference in slopes</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>??</td>
<td>1</td>
<td>0.0094 (0.0086)</td>
<td>0.27</td>
</tr>
<tr>
<td>Power (Autoregressive)</td>
<td>??</td>
<td>2</td>
<td>0.0094 (0.0032)</td>
<td>0.0038</td>
</tr>
<tr>
<td>Random Regression</td>
<td>??</td>
<td>4</td>
<td>0.0089 (0.0031)</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

- Confidence intervals and tests depend on the covariance!

Outline

Random regression

General repeated measurements

- Random effects ANOVA (the two-level model)
- Variance components in general
- Zero estimates of variance components
- Linear mixed models (LMMs)
- Comparing measurement methods

What are repeated measurements?

Repeted measurements refer to data where the same outcome has been measured in different situations (or at different spots) on the same individuals.

- Special case: longitudinal means repeatedly over time.

What is clustered data?

Repeated measurements are termed clustered data when the same outcome is measured on groups of individuals from the same families/workplaces/school classes/villages/etc.
Analysis of repeated measurements

Many applications:
- Longitudinal data
- Treatments applied to multiple limbs, teeth, etc within the same person.
- Cross-over trials.
- Cluster randomized trials / multi-center studies.
- Comparisons / reliability of measurement methods.

ATT: Measurements belonging to the same subject/cluster are correlated. If we fail to take this correlation into account we will experience:
- p-values that are too small or too large.
- confidence intervals that are too wide or too narrow.

Sources of variation / correlation

Measurements belonging to the same subject/cluster tend to be correlated (look alike) due to e.g.
- Environmental variation.
  - Between regions, hospitals or countries.
- Biological variation.
  - Between individuals, families or animals.

Today: Use random effects (variance components) to model various sources of variation in a linear mixed model framework.

Outline

Random regression
General repeated measurements
Random effects ANOVA (the two-level model)
Variance components in general
Zero estimates of variance components
Linear mixed models (LMMs)
Comparing measurement methods

One-way analysis of variance – with random variation

Comparison of \( k \) groups or clusters, satisfying:
- The groups are of no individual interest and it is of no relevance to test whether they have identical means.
- The groups may be thought of as representatives from a population, that we want to describe.
Example: Rabbit data

- $R = 6$ rabbits vaccinated.
- In $S = 6$ spots on the back.

Response: swelling in cm$^2$

Research question: How much swelling can be expected in reaction to the vaccine?

Random effects anova (the two-level model)

We let each rabbit have its own level of swelling described as

$$Y_{rs} = A_r + \varepsilon_{rs}$$

- We assume that these individual levels are randomly sampled from a normally distributed population,

$$A_r \sim \mathcal{N}(\mu, \omega_B^2)$$

- The error terms are considered to be independent normal,

$$\varepsilon_{rs} \sim \mathcal{N}(0, \sigma_W^2)$$

The rabbit levels are so-called random effects and the variances $\omega_B^2$ and $\sigma_W^2$ are so-called variance components describing the variance between rabbits and within rabbits, respectively.

Implications of random effects anova

All observations are considered as randomly sampled measurements from the same population. Thus, the model implies that all measurements follow the same normal distribution:

$$Y_{rs} \sim \mathcal{N}(\mu, \omega_B^2 + \sigma_W^2)$$

- Population mean $\mu$, the grand mean.
- Population variance $\omega_B^2 + \sigma_W^2$, the total variation.

But: Measurements made on the same rabbit are correlated with the so-called intra-class correlation

$$\text{Corr}(y_{r1}, y_{r2}) = \rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2}$$

Compound symmetry

The implied covariance of the repeated measurements has a compound symmetry-structure:

$$
\begin{pmatrix}
\omega_B^2 + \sigma_W^2 & \omega_B^2 & \ldots & \omega_B^2 \\
\omega_B^2 & \omega_B^2 + \sigma_W^2 & \ldots & \omega_B^2 \\
\vdots & \vdots & \ddots & \vdots \\
\omega_B^2 & \omega_B^2 & \ldots & \omega_B^2 + \sigma_W^2
\end{pmatrix}
$$

In particular all pairs of spots on the same rabbit are assumed to be equally correlated (with the intra-class correlation).

- We say that the spots are exchangeable.

Note: If this is not the case, an unstructured covariance might fit the data better. Say, if some spots are expected to respond more similarly than others.
Random effects ANOVA in PROC MIXED

```plaintext
PROC MIXED DATA=rabbit;
  CLASS rabbit;
  MODEL swelling = / SOLUTION;
  RANDOM rabbit;
RUN;
```

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>rabbit</td>
<td>0.3304</td>
</tr>
<tr>
<td>Residual</td>
<td>0.5842</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect   | Estimate | Error  | DF  | t Value | Pr > |t| |
|----------|----------|--------|-----|---------|-------|---|
| Intercept| 7.3667   | 0.2670 | 5   | 27.59   | <.0001|

Estimation of variance components

<table>
<thead>
<tr>
<th>Level</th>
<th>Variation</th>
<th>Variance component</th>
<th>Estimate</th>
<th>% of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Between</td>
<td>( \omega^2_B )</td>
<td>0.3304</td>
<td>36%</td>
</tr>
<tr>
<td>2</td>
<td>Within</td>
<td>( \omega^2_W )</td>
<td>0.5842</td>
<td>64%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>( \omega^2_B + \sigma^2_W )</td>
<td>0.9146</td>
<td>100%</td>
</tr>
</tbody>
</table>

We can use the covtest-option in

```plaintext
PROC MIXED COVTEST DATA=rabbit; ...
```

to get standard errors for the variance components:

- 95%CI for **Intra**-rabbit variation \( \sigma^2_W \): (0.37 – 1.04).
- 95%CI for **Inter**-rabbit variation \( \omega^2_B \): (0.06 – 2.48).

Beware not to overinterpret the estimates in a small dataset!

Computing variance components

In balanced data (same number of observations per cluster):

**Explicit solution:**

\[
\tilde{\sigma}^2_W = MS_W \quad \text{and} \quad \tilde{\omega}^2_B = \frac{MS_B - MS_W}{N}
\]

- \( N \) is the number of clusters
- \( MS_W \) and \( MS_B \) are Mean Squares within and between clusters, defined as in one-way ANOVA.

This is deduced from \( E(MS_B) = N\omega^2_B + \sigma^2_W \) and \( E(MS_W) = \sigma^2_W \).

Typical differences

Difference between spots on the **same** rabbit:

\[
y_{r_1s_1} - y_{r_2s_2} = \varepsilon_{r_1s_1} - \varepsilon_{r_2s_2} \\
\sim N(0, 2\omega^2_W)
\]

- **Normal region:** \( \pm 2 \sqrt{2\omega^2_W} = \pm 2.16 \text{ cm}^2 \)

Difference between spots on **different** rabbits:

\[
y_{r_1s_1} - y_{r_2s_2} = \alpha_{r_1} - \alpha_{r_2} + \varepsilon_{r_1s_1} - \varepsilon_{r_2s_2} \\
\sim N(0, 2\sigma^2_B + 2\omega^2_W)
\]

- **Normal region:** \( \pm 2 \sqrt{2\sigma^2_B + 2\omega^2_W} = \pm 2.70 \text{ cm}^2 \)
Why not use traditional one-way anova?

Focus on rabbit means: and test \( H_0: \mu_1 = \ldots = \mu_6. \)

One-way anova table:

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS = SS/df</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between rabbits</td>
<td>12.8333</td>
<td>( R - 1 ) = 5</td>
<td>2.5667</td>
<td>4.39</td>
</tr>
<tr>
<td>Within rabbit</td>
<td>17.5266</td>
<td>( R(S - 1) ) = 30</td>
<td>0.5842</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30.3599</td>
<td>( RS - 1 ) = 35</td>
<td>0.8674</td>
<td></td>
</tr>
</tbody>
</table>

Test for identical rabbits means: \( F = 4.39 \sim F(5, 30), \) \( P = 0.004. \)

But: We are not interested in these particular 6 rabbits, only in rabbits in general, as a species! Presumably these 6 rabbits have been randomly sampled from the species.

One-way anova with and without random variation

Classical one-way anova

- The rabbit means \( \mu_r \) are fixed parameters, - supposedly of an interest of their own.
- We say that the rabbit factor is a fixed effect.

Random effects one-way anova

- The rabbit levels \( A_r \) are considered random and their population mean \( \mu \) and variance \( \omega^2_B + \sigma^2_W \) is the major interest.
- We say that the rabbit factor is a random effect.
- (If data is from a pilot study used in the planning of some trial, the intra-class correlation will also be of interest).

Comparison of modeling strategies

Quantifying overall swelling

Four strategies for estimating the grand mean (i.e. of the rabbit population).

<table>
<thead>
<tr>
<th>method</th>
<th>estimate (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: forget rabbit</td>
<td>7.367 (0.155)</td>
</tr>
<tr>
<td>2: fixed rabbit</td>
<td>7.367 (0.127)</td>
</tr>
<tr>
<td>3: rabbit averages</td>
<td>7.367 (0.267)</td>
</tr>
<tr>
<td>4: random rabbit</td>
<td>7.367 (0.267)</td>
</tr>
</tbody>
</table>

1. We assume independence between all 36 measurements
2. We estimate the mean swelling of exactly these 6 rabbits by classical one-way anova
3. We analyse the sample of averages for the six rabbits (summary statistics).
4. We estimate the mean swelling of rabbits as a species in the random effects anova model (the correct approach)

Comments on the strategies:

1. Ignoring the clustering is wrong!
   - leads to systematic underestimation of the standard error.
2. In the fixed effect one-way anova the grand mean has a different interpretation!
   - leads to systematic underestimation of the standard error.
3. Looking at the sample of averages may be OK.
   - At least in balanced designs (otherwise the individual averages have unequal variances and the standard error may be affected)
   - But we lose all information on within subject variation.
   (E.g. not possible to test for systematic spot-differences.)
Comparison of modeling strategies

When the 3 smallest measurements from rabbit 2 (largest level) are omitted, the results become:

<table>
<thead>
<tr>
<th>method</th>
<th>estimate (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: forget rabbit</td>
<td>7.291 (0.163)</td>
</tr>
<tr>
<td>2: fixed rabbit</td>
<td>7.291 (0.136)</td>
</tr>
<tr>
<td>3a: rabbit averages</td>
<td>7.291 (0.265)</td>
</tr>
<tr>
<td>(weighted)</td>
<td></td>
</tr>
<tr>
<td>3b: rabbit averages</td>
<td>7.436 (0.333)</td>
</tr>
<tr>
<td>(unweighted)</td>
<td></td>
</tr>
<tr>
<td>4: random rabbit</td>
<td>7.390 (0.298)</td>
</tr>
<tr>
<td>Full sample</td>
<td>7.367 (0.267)</td>
</tr>
</tbody>
</table>

1 we have omitted some of the largest observations
2+3a rabbit 2 has a lower weight in the average (only 3 observations)
3b average for rabbit 2 has increased
4 rabbit 2 has a lower weight in the average due to a larger standard error

Estimation of individual rabbit means

Sometimes estimates of individual random effects are used for e.g. prediction of future disease status.

How do we estimate them?

- Simple averages $\bar{y}_r$ of the individual measurements.
- Best unbiased linear predictors (BLUPs) are weighted averages of the individual and the population mean:

$$\hat{y}_r = \frac{\omega^2_B}{\omega^2_B + \frac{\sigma^2_w}{S}} \bar{y}_r + \frac{\frac{\sigma^2_w}{S}}{\omega^2_B + \frac{\sigma^2_w}{S}} \bar{y}_r.$$

They have been shrunked towards the grand mean, $\bar{y}_r$.

BLUPs vs averages

Full data

Reduced data

Note: We see larger shrinkage for rabbit no. 2 when the 3 smallest measurements from this rabbit have been removed (i.e. we are borrowing strength from the neighbours).

Outline

- Random regression
- General repeated measurements
- Random effects ANOVA (the two-level model)
- Variance components in general
- Zero estimates of variance components
- Linear mixed models (LMMs)
- Comparing measurement methods
General variance component models

Generalisations of ANOVA and GLM models involving several sources of random variation, so-called variance components.

Examples of sources of random variation:
- Environmental variation.
  - Between regions, hospitals or countries.
- Biological variation.
  - Between individuals, families or animals.
- Within-individual variation.
  - Between arms, teeth, days.
- Variation due to uncontrollable circumstances.
  - E.g. time of day, temperature, observer.
- Measurement error.

Multilevel models

Variance component models are also called multilevel models.

- Levels are most often hierarchical.
- We have variation, i.e. a variance component, on each level.
- And possibly systematic effects (covariates) on each level.

Example: A three-level model

Outcome: Number of nuclei per cell in the rat pancreas (used for the evaluation of cytostatica)

- $R = 4$ rats.
- $S = 3$ sections for each rat.
- $F = 5$ randomly chosen fields from each section.

<table>
<thead>
<tr>
<th>level 1 → level 2 → level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>fields → sections → rats</td>
</tr>
<tr>
<td>$\sigma^2$ $\tau^2$ $\omega^2$</td>
</tr>
</tbody>
</table>

Typical differences (normal regions)

- For sections on **different rats**:
  \[ \pm 2 \times \sqrt{2} \times (0.0179 + 0.0029 + 0.1968) = \pm 1.319 \]

- For **different sections** on the **same rat**:
  \[ \pm 2 \times \sqrt{2} \times (0.0029 + 0.1968) = \pm 1.264 \]

- For **different fields** on the **same section**:
  \[ \pm 2 \times \sqrt{0.1968} = \pm 1.255 \]

Correlation

Estimated correlations between two measurements on the same rat:

- If they are measured on the **same section**:
  \[ \text{Corr}(y_{rs1}, y_{rs2}) = \frac{\omega^2 + \tau^2}{\omega^2 + \tau^2 + \sigma^2} = 0.096. \]

- If they are measured on **different sections**:
  \[ \text{Corr}(y_{r11}, y_{r22}) = \frac{\omega^2}{\omega^2 + \tau^2 + \sigma^2} = 0.082. \]

Merits of multilevel models

We get a **better understanding** of the various sources of variation.

Effects **within** may be **estimated more precisely** (higher power), since some sources of variation are eliminated, e.g. by making comparisons within a family. This is analogous to the **paired comparison** situation.

When planning investigations, estimates of the variance components are needed in order to compare the power of various designs, and help us decide

- How many replicates do we need at each level?
- Should we randomize entire clusters or randomize **within** the clusters?

Design considerations

(Nota**ne in analogy with cluster-randomized trials.**)

Plan an experiment with:

- **R** rabbits.
- **S** spots for each rabbit.
- **R × S** measurements.

Std. error of grand mean,

\[ \text{var}(\bar{y}) = \frac{\omega^2}{R} + \frac{\sigma^2}{RS}, \]

decreases with **R** and **S**.

The different curves correspond to **S** varying from 1 to 10.
**Effective sample size**

How many rabbits would we need to obtain the same precision in estimating the grand mean if we had only one measurement on each of \( R_1 \) rabbits?

Solve an equation to get:

\[
R_1 = \frac{R \times S}{1 + \rho(S - 1)}
\]

where \( \rho \) is the within rabbit correlation.

- Estimate: \( \rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2} = \frac{0.3304}{0.3304 + 0.5842} = 0.361 \Rightarrow R_1 = 12.8 \)

I.e. one measurement on each of thirteen rabbits gives the same precision as six measurements on each of six rabbits.

**Drawbacks of multilevel models**

Their statistical analysis is more difficult.

- When making inference (estimation and testing), it is important to take all sources of variation into account, and effects have to be evaluated against the relevant variation.

If we fail to take the correlation into account, we will experience:

- Possible bias in the mean value estimates.
- Too small standard errors (type 1 error) for estimates of level 2 covariates (between-cluster effects).
- Too large standard errors (type 2 error) for estimates of level 1 covariates (within-cluster effects)

**Fixed or random effect?**

How do we decide whether a factor should be modeled as fixed or random?

**Fixed**

- The specific values of the factor have been predetermined when planning the study.
- Allows inference for these particular values only.
- Demands a decent number of observations in each group.

**Random**

- A representative sample of values of the factor is present.
- Allows inference to be extended beyond the values in the experiment and to the population they were sampled from.

**Outline**

- Random regression
- General repeated measurements
- Random effects ANOVA (the two-level model)
- Variance components in general
- Zero estimates of variance components
- Linear mixed models (LMMs)
- Comparing measurement methods
Example: Cortisol

Outcome: Concentration of cortisol in blood samples taken morning and evening in workers in Aarhus amt and kommune in 2007 (3536 participants) with similar follow-up in 2009 (2408 participants)

Interest: effect of stressors: lifeevents, Effort Reward Index variation
between persons gender, age within person: between days bmi, stressors within person: within days time (of day)

Sample data
From 8 randomly selected men:

NOTE: concentrations on logarithmic scale.

Multi-level analysis
PROC MIXED DATA=prism COVTEST; WHERE sex EQ 'male';
CLASS id year time;
MODEL logcortisol = time / SOLUTION CL DDFM=SATTERTH;
RANDOM id id*year;
RUN;

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Std.Error</th>
<th>Z Value</th>
<th>Pr &gt; Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>0.05993</td>
<td>0.01266</td>
<td>4.73</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>id*year</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>0.5385</td>
<td>0.01794</td>
<td>30.01</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Run</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>1305</td>
<td>4916.89</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Negative variance components
In case one of the variance component estimates becomes negative, SAS reports a zero.

What does it mean?
▶ The zero-estimate may be a chance finding due to statistical uncertainty.
▶ Or it might be the result of truly negative correlation within clusters - e.g. from competition (plants grown in same pot).

What can we do about it?
▶ Re-fit the model without the problematic random effect.
▶ Use a covariance pattern model which allows for negative correlation.
▶ Include more covariates at the lower levels.
Estimated variance components

<table>
<thead>
<tr>
<th>Level</th>
<th>Variation</th>
<th>Estimate</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>between persons ($\omega^2$)</td>
<td>0.0599</td>
<td>10.0%</td>
</tr>
<tr>
<td>2</td>
<td>between days ($\tau^2$)</td>
<td>0.0000</td>
<td>0.0%</td>
</tr>
<tr>
<td>1</td>
<td>within days ($\sigma^2$)</td>
<td>0.5385</td>
<td>90.0%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.5984</td>
<td>100%</td>
</tr>
</tbody>
</table>

Level 2 covariates (stressors) can only have very little impact on individual cortisol concentrations!

Systematic effects

Cortisol is measured on log-scale. Backtransformation $\exp(2.0137) \simeq 7.49$ yields that median levels of kortisol is an estimated 7.5 times higher in the morning than in the evening.

Exact time of measurement should be taken into account!!!

Outline

- Random regression
- General repeated measurements
- Random effects ANOVA (the two-level model)
- Variance components in general
- Zero estimates of variance components
- Linear mixed models (LMMs)
- Comparing measurement methods

Specification of linear mixed models (LMMs)

(the "M" in the middle refers to mixed fixed and random effects)

Systematic variation

- covariates: time, treatment, gender, age, etc.,
  describing population parameters.

Random variation:

- Random effects,
  describing subject specific parameters.
- Serial correlation
- Measurement error

Note: Interactions between systematic and random effects are always random effects.
Technical model description for LMMs
(the surrounding "L-M" refers to a linear model specification)

Model repeated outcomes on subject/cluster \( i \) as:

\[
Y_i = X_i \beta + Z_i b_i + \varepsilon_i
\]

using vector- and matrix-notation . . . sorry!

- **Systematic effects** \( \beta \) with designmatrices \( X_i \).
- **Random effects** \( b_i \) with designmatrices \( Z_i \).
- Possibly dependent **residual error terms** \( \varepsilon_i \).

We assume that the \( b_i \)'s and \( \varepsilon_i \)'s are independent normally distributed with mean zero and covariance matrices given by:

- **The G-matrix**: \( \text{Var}(b_i) = G \).
- **The R-matrix**: \( \text{Var}(\varepsilon_i) = R \).

Implied covariance for LMMs

The covariance of the repeated measurements on subject/cluster \( i \) is given by the general formula:

\[
V_i = Z_i^T G Z_i + R_i
\]

. . . using matrix-multiplication (sorry again!)

Note:
- This is the so-called **V-matrix**.
- Print with option `vcorr` in `proc mixed`.

SAS: PROC MIXED

- **model**: describes the mean value structure (i.e. covariates / fixed effects)
- **random**: describes the random effects
- **repeated**: describes the residual covariance

Very flexible modeling framework!

Example: It is possible to model
- longitudinal series of measurements (level 1) . . .
- with repeated series on each subject and with different treatments along the way (level 2) . . .
- and subjects belonging to different clusters (level 3).

Nonidentifiability

**Warning**: Make sure you understand your model!

- Modeling random effects together with a residual error covariance may result in unidentifiable covariance parameters, i.e. nonconvergence, unless done with some care.

Example: Compound symmetry can be specified as either of:
- random intercept / subject=id;
- repeated time / subject=id type=cs;

in case both lines are included in the **same program**, it will not converge.
Comparing measurement devices


Example: Peak expiratory flow rate, l/min:

- 17 subjects, 2 measurement devices,
  - each measured twice

<table>
<thead>
<tr>
<th>subject</th>
<th>Wright</th>
<th>mini Wright</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Y₁₁</td>
<td>Y₁₂</td>
</tr>
<tr>
<td>1</td>
<td>494</td>
<td>490</td>
</tr>
<tr>
<td>2</td>
<td>395</td>
<td>397</td>
</tr>
<tr>
<td>3</td>
<td>516</td>
<td>512</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>178</td>
<td>165</td>
</tr>
<tr>
<td>16</td>
<td>423</td>
<td>372</td>
</tr>
<tr>
<td>17</td>
<td>427</td>
<td>421</td>
</tr>
</tbody>
</table>

Average 450.35 445.41 452.47 455.35
SD 116.31 119.61 113.12 111.32

Aim of investigation

Quantify the precision of each measuring device
  - Variability.
  - Reproducibility.

Quantify the agreement between the two devices
  - Bias of one method compared to the other.
  - Variance of one method compared to the other.

Can the devices be used interchangeably in clinic?
Simple approaches

For reliability

▷ Compare the replicate measurements in Bland-Altman plots*.
▷ Each method separately.

For method comparison

▷ Compare averages in a Bland-Altman plot?
▷ Not good - unless you also do averages in clinic!

⋆ See: Bland & Altman (1986) for further explanation.

Variance component model

<table>
<thead>
<tr>
<th>level</th>
<th>variation</th>
<th>covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>between subjects ($\omega^2$)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>between methods ($\tau^2$)</td>
<td>method</td>
</tr>
<tr>
<td>1</td>
<td>within methods ($\sigma^2$)</td>
<td></td>
</tr>
</tbody>
</table>

Specified as:

\[ Y_{ijk} = \mu_j + A_i + B_{ij} + \varepsilon_{ijk} \]

▷ $A_i \sim N(0, \omega^2)$ for subjects $i = 1, \ldots, 17$,
▷ $B_{ij} \sim N(0, \tau^2)$ for methods $j = 1, 2$,
▷ $\varepsilon_{ijk} \sim N(0, \sigma^2)$ for replicate $k = 1, 2$.

Assuming same residual variance for both methods ...

Covariance structure

▷ We have 4 measurements on each subject

Covariance matrix with ordering (wright1, wright2, mini1, mini2):

\[
\begin{pmatrix}
\omega^2 + \tau^2 + \sigma^2 & \omega^2 + \tau^2 & \omega^2 & \omega^2 \\
\omega^2 + \tau^2 & \omega^2 + \tau^2 + \sigma^2 & \omega^2 & \omega^2 \\
\omega^2 & \omega^2 & \omega^2 + \tau^2 + \sigma^2 & \omega^2 + \tau^2 \\
\omega^2 & \omega^2 & \omega^2 + \tau^2 & \omega^2 + \tau^2 + \sigma^2
\end{pmatrix}
\]

▷ We expect stronger correlation between measurements made with the same method than with different methods.

Analysis

PROC MIXED DATA=wright;
  CLASS method id;
  MODEL flow=method / SOLUTION CL;
  RANDOM intercept method / SUBJECT=id;
RUN;

| Effect | method | Estimate | Standard Error | DF  | t Value | Pr > |t| |
|--------|--------|----------|----------------|-----|---------|-----|-----|
| Intercept | 447.88 | 27.7519  | 16  | 16.14 | < .0001 |
| method   | mini   | 6.0294   | 8.0532        | 16  | 0.75    | 0.4649 |
| method   | wright | 0        | .              | .   | .       | .    |

Solution for Fixed Effects

No evidence of systematic differences between the measurement methods.
Estimated variance components

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>id</td>
<td>12542</td>
</tr>
<tr>
<td>method</td>
<td>id</td>
<td>393.57</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>315.37</td>
</tr>
</tbody>
</table>

Fit Statistics
-2 Res Log Likelihood: 676.0
AIC (smaller is better): 681.6

What does this tell us about the precision of the measurements?

Typical differences

Between replicate measurements using the same method:

\[ Y_{ijk1} - Y_{ijk2} = \varepsilon_{ijk1} - \varepsilon_{ijk2} \sim N(0, 2\sigma^2) \]

Limits-of-agreement: \( \pm 2\sqrt{2\sigma^2} \simeq \pm 50.23 \).

Between measurements using different methods:

\[ Y_{ij1k1} - Y_{ij2k1} = \mu_j - \mu_j + B_{ij1} - B_{ij2} + \varepsilon_{ij1k1} - \varepsilon_{ij2k1} \sim N(\mu_j - \mu_j, 2\tau^2 + 2\sigma^2) \]

Limits-of-agreement: \( \mu_1 - \mu_2 \pm 2\sqrt{2\tau^2 + 2\sigma^2} \simeq 6.03 \pm 75.31 \).

(where we include the non-significant systematic difference).

Comparing precisions

We need a slightly more general model:

\[ Y_{ijk} = \mu_j + A_i + B_{ij} + \varepsilon_{ijk} \]

\[ A_i \sim N(0, \omega^2) \text{ for subjects } i = 1, \ldots, 17, \]
\[ B_{ij} \sim N(0, \tau^2) \text{ for methods } j = 1, 2, \]
\[ \varepsilon_{ijk} \sim N(0, \sigma_j^2) \text{ for replicate } k = 1, 2. \]

Now: method-dependent residual variance-

Analysis

PROC MIXED DATA=wright;
  CLASS method id;
  MODEL flow=method / SOLUTION CL DDFM=SATTERTHWAIT;
  RANDOM intercept method / SUBJECT=id;
  REPEATED / GROUP=method type=SIMPLE SUBJECT=id*method;
RUN;

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Group</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>id</td>
<td></td>
<td>12542</td>
</tr>
<tr>
<td>method</td>
<td>id</td>
<td></td>
<td>393.57</td>
</tr>
<tr>
<td>Residual</td>
<td>methodid</td>
<td>method mini</td>
<td>396.44</td>
</tr>
<tr>
<td>Residual</td>
<td>methodid</td>
<td>method wright</td>
<td>234.29</td>
</tr>
</tbody>
</table>

Fit Statistics
-2 Res Log Likelihood: 474.8
AIC (smaller is better): 480.8
Comparing precisions

Estimated precisions:

Wright: \( \hat{\sigma}^2_1 = 234.29 \rightarrow \pm 2\sqrt{2\sigma^2_1} \simeq XXXX \)

mini: \( \hat{\sigma}^2_1 = 396.44 \rightarrow \pm 2\sqrt{2\sigma^2_1} \simeq XXXX \)

Seemingly Wright is more precise, but is the difference significant?

\[-2 \log Q = 676.0 - 674.8 = 1.2 \sim \chi^2(1) \rightarrow P = 0.27\]

Don’t form too firm a conclusion with too small data.

Overall comparison

Solution for Fixed Effects

| Effect       | method | Estimate | Error  | DF | t Value | Pr > |t| |
|--------------|--------|----------|--------|----|---------|------|---|
| Intercept    |        | 447.88   | 27.7519| 16 | 16.14   | <.0001 |
| method       | mini   | 6.0294   | 8.0532 | 16 | 0.75    | 0.4649 |
| method       | wright | 0        | .      | .  | .       | .    |

No evidence of systematic differences between the measurement methods.

Typical differences between the two methods:

\[ Y_{ij_1k_1} - Y_{ij_2k_1} = \mu_{j_1} - \mu_{j_2} + B_{ij_1} - B_{ij_2} + \epsilon_{ij_1k_1} - \epsilon_{ij_2k_1} \sim \mathcal{N}(\mu_{j_1} - \mu_{j_2}, 2\tau^2 + \sigma^2_1 + \sigma^2_2) \]

Limits-of-agreement:

\[ \mu_1 - \mu_2 \pm 2\sqrt{2\tau^2 + \sigma^2_1 + \sigma^2_2} \simeq 6.03 \pm XXXX. \]