Variance component models
Analysis of repeated measurements, NFA 2016

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

Linear mixed models for clustered data and repeated measurements in general, i.e. not just for longitudinal data.

New concepts:

▶ random effects
▶ variance components
▶ multi-level models

Suggested reading:

▶ Fitzmaurice et al. (2011): chapters 8, 21, 22.
Outline

General repeated measurements

Random effects ANOVA (the two-level model)

Fixed vs random effects

Multi-level models

Ecological fallacy

Comparing measurement methods
Analysis of repeated measurements

Many applications:

- Longitudinal data (lecture 2)
- Cluster randomized trials/multi-center studies.
- Reproducibility/reliability of measurement methods.
- Treatments applied to multiple limbs, teeth, etc within the same subject.
- Cross-over trials (lecture 4).

**ATT:** Measurements belonging to the same subject/cluster are correlated. If we fail to take correlation into account our statistical results may be biased.
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 work places.

- 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

General repeated measurements

Random effects ANOVA (the two-level model)

Fixed vs random effects

Multi-level models

Ecological fallacy

Comparing measurement methods
One-way analysis of variance – with random variation

The simplest possible model for clustered data.

- 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:

\[ Y_{rs} \sim N(\mu, \omega^2_B + \sigma^2_W) \]

- Population mean \( \mu \), the grand mean.
- Population variance \( \omega^2_B + \sigma^2_W \), the total variation.

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

\[ \text{Corr}(y_{r1}, y_{r2}) = \rho = \frac{\omega^2_B}{\omega^2_B + \sigma^2_W} \]
Compound symmetry

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

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

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

If any two pairs of measurements are equally correlated we say that the measurements are exchangeable.

▶ Are the spots randomly selected - ???

If not, an unstructured covariance is more appropriate

▶ Some spots are expected to respond more similarly than others (physiological/spatial correlation pattern).

In other situations exchangeability is obvious

▶ E.g. patients sampled randomly from several GPs.
Random effects anova in PROC MIXED

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>

\[ ICC = \frac{\omega^2_B}{\omega^2_B + \sigma^2_W} = 0.36. \]

Quite a lot of variability **within** rabbits - ?

- Are there systematic differences between the spots?
- Or perhaps measurements just aren’t that precise.

Beware not to **overinterpret** the estimates in a small dataset!
Interpreation of variance components

Typical differences between spots on **different** rabbits:

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

- **95% normal range**: \(0 \pm 2\sqrt{2\sigma^2_B + 2\omega^2_W} = \pm 2.70 \ cm^2\)

Typical differences between spots on the **same** rabbit:

\[ y_{rs_1} - y_{rs_2} = \varepsilon_{rs_1} - \varepsilon_{rs_2} \sim N(0, 2\omega^2_W) \]

- **95% normal range**: \(0 \pm 2\sqrt{2\omega^2_W} = \pm 2.16 \ cm^2\)
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 $\omega_B^2 + \sigma_W^2$ 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).
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:

$$\tilde{\omega}_B^2 \bar{y}_r + \frac{\tilde{\sigma}_W^2}{S} \bar{y}. + \frac{\tilde{\omega}_B^2}{\tilde{\omega}_B^2 + \frac{\tilde{\sigma}_W^2}{S}} \bar{y}..$$

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

General repeated measurements

Random effects ANOVA (the two-level model)

Fixed vs random effects

Multi-level models

Ecological fallacy

Comparing measurement methods
Fixed or random effect?

Fixed effects such as treatment, gender, and time.
- Typically a limited number of carefully selected groups.
- Group names are specific and cannot be shuffled.
- Each group must have a decent size in order to reach interesting conclusions (statistical power).

Random effect such as subject, rat or family.
- Possibly a large number of different groups.
- Group names are non-informative (number of subject, rat or family) and could be shuffled without consequence.
- Allows inference to be extended beyond the subjects in the experiment and to the population they were sampled from.
- The number of groups matters not the size of the groups.
Testing fixed effects

Imagine that rabbits are grouped in two (e.g. treatments):

<table>
<thead>
<tr>
<th>level</th>
<th>variation</th>
<th>covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>within rabbit</td>
<td>spot</td>
</tr>
<tr>
<td>2</td>
<td>between rabbits</td>
<td>group</td>
</tr>
</tbody>
</table>

- Part of the variation *between rabbits* could be explained by systematic differences between groups.
- Part of the variation *within rabbits* could be explained by systematic differences between spots.
Testing fixed effects with PROC MIXED

PROC MIXED DATA=.rabbit;
   CLASS group rabbit spot;
   MODEL swelling = group spot / SOLUTION CL DDFM=KR;
   RANDOM rabbit;
RUN;

Output:

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>smaller than before</th>
</tr>
</thead>
<tbody>
<tr>
<td>rabbit</td>
<td>0.3694</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>0.5477</td>
<td></td>
</tr>
</tbody>
</table>
Testing fixed effects with PROC MIXED

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>1</td>
<td>4</td>
<td>0.64</td>
<td>0.4675</td>
</tr>
<tr>
<td>spot</td>
<td>5</td>
<td>25</td>
<td>1.40</td>
<td>0.2584</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>spot</th>
<th>group</th>
<th>Estimate</th>
<th>StdError</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.9111</td>
<td>0.4792</td>
<td>4</td>
<td>14.42</td>
<td>0.0001</td>
<td>0.05</td>
<td>5.5807</td>
<td>8.2416</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group</td>
<td>1</td>
<td>0.4444</td>
<td>4</td>
<td>0.80</td>
<td>0.4675</td>
<td>0.05</td>
<td>-1.0942</td>
<td>1.9831</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>group</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1.52</td>
<td>0.1408</td>
<td>0.05</td>
<td>-0.2300</td>
<td>1.5300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spot</td>
<td>a</td>
<td>0.6500</td>
<td>25</td>
<td>0.12</td>
<td>0.9078</td>
<td>0.05</td>
<td>-0.8300</td>
<td>0.9300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spot</td>
<td>b</td>
<td>0.0500</td>
<td>25</td>
<td>0.12</td>
<td>0.9078</td>
<td>0.05</td>
<td>-0.8300</td>
<td>0.9300</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
Disregarding repeated measurements

When the *random rabbit variation* is ignored:

```plaintext
PROC GLM DATA=rabbit;
  CLASS group spot;
  MODEL swelling=group spot / SOLUTION CLPARM;
RUN;
```

### Analysis Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>1</td>
<td>1.77777778</td>
<td>1.77777778</td>
<td>2.08</td>
<td>0.1596</td>
</tr>
<tr>
<td>spot</td>
<td>5</td>
<td>3.83333333</td>
<td>0.76666667</td>
<td>0.90</td>
<td>0.4954</td>
</tr>
</tbody>
</table>

| Parameter       | Estimate | Standard Error | t Value | Pr > |t| | 95% Confidence Limits |
|-----------------|----------|----------------|---------|------|---|-----------------------|
| Intercept       | 6.911111111 B | 0.40735835 | 16.97   | < |0.0001 | 6.077969737 7.744252485 |
| group 1         | 0.444444444 B | 0.30793397 | 1.44    | 0.1596 | -0.185351236 1.074240125 |
| group 2         | 0.000000000 B | .          | .       | .     | .                      |
| spot a          | 0.650000000 B | 0.53335728 | 1.22    | 0.2328 | -0.440838117 1.740838117 |
| spot b          | 0.050000000 B | 0.53335728 | 0.09    | 0.9260 | -1.040838117 1.140838117 |
| ...             |          |              |         |       |                        |

*Too small standard errors* for estimates of difference between groups and *too large standard errors* for estimates of differences between spots!
Outline

General repeated measurements

Random effects ANOVA (the two-level model)

Fixed vs random effects

Multi-level models

Ecological fallacy

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 work places.
- 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.

<table>
<thead>
<tr>
<th>individual</th>
<th>→</th>
<th>context/cluster</th>
<th>→</th>
<th>context/cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>level 1</td>
<td>→</td>
<td>level 2</td>
<td>→</td>
<td>level 3</td>
</tr>
<tr>
<td>students</td>
<td>→</td>
<td>classes</td>
<td>→</td>
<td>schools</td>
</tr>
<tr>
<td>patient</td>
<td>→</td>
<td>clinic</td>
<td>→</td>
<td>regions</td>
</tr>
<tr>
<td>visit</td>
<td>→</td>
<td>girl</td>
<td>→</td>
<td></td>
</tr>
<tr>
<td>spot</td>
<td>→</td>
<td>rabbit</td>
<td>→</td>
<td></td>
</tr>
</tbody>
</table>
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

(Note in analogy with cluster-randomized trials.)

Plan an experiment with:

- $R$ rabbits.
- $S$ spots for each rabbit.
- $R \times S$ measurements.

Std. error of grand mean,

$$\text{var}(\bar{y}) = \frac{\omega_B^2}{R} + \frac{\sigma_W^2}{RS},$$

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).
Case: Cortisol and stress-response

Outcome: Concentration of cortisol in salvia samples taken mornings and evenings in workers in Aarhus amt and kommune in 2007 (3536 participants, 786 men) with follow-up in 2009 (2408 participants, 520 men).

Interest: effect of stressors: life events, Effort Reward Index.

<table>
<thead>
<tr>
<th>level</th>
<th>variation</th>
<th>covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>between persons</td>
<td>gender, age</td>
</tr>
<tr>
<td>2</td>
<td>within person: between days</td>
<td>bmi, stressors, year</td>
</tr>
<tr>
<td>1</td>
<td>within person: within days</td>
<td>timeday (morning/evening)</td>
</tr>
</tbody>
</table>

Reference: from PRISM study, personal communication with Sigurd Mikkelsen.
Log-transformed concentrations
Three-level model

title1 'variance components';
PROC MIXED DATA=prism_men;
   CLASS idnr year (ref='2007') timeday;
   MODEL logkonc = timeday year / SOLUTION CL DDFM=KR;
   RANDOM idnr idnr*year;
RUN;

Iteration Evaluations -2 Res Log Like Criterion
0 1 6077.88355058
1 3 6050.14347396 0.00008342
2 1 6050.09026809 0.00000005
3 1 6050.09023526 0.00000000

Convergence criteria met.

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>idnr</td>
<td>0.05943</td>
</tr>
<tr>
<td>idnr*year</td>
<td>0</td>
</tr>
<tr>
<td>Residual</td>
<td>0.5374</td>
</tr>
</tbody>
</table>

One of the variance component estimates is a zero!
Estimated variance components

<table>
<thead>
<tr>
<th>Level</th>
<th>Variation</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>between persons ($\omega^2$)</td>
<td>0.0594 (10.0%)</td>
</tr>
<tr>
<td>2</td>
<td>between days ($\tau^2$)</td>
<td>0.0000 (0.0%)</td>
</tr>
<tr>
<td>1</td>
<td>within days ($\sigma^2$)</td>
<td>0.5374 (90.0%)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0.5984 (100%)</td>
</tr>
</tbody>
</table>

Level 2 covariates (stressors) can only have very little impact on individual cortisol concentrations!
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. competition between plants grown in same pot.

What can we do about it?

- Re-fit the model without the problematic random effect.
- Use an unstructured covariance allowing negative correlation
- Include more level 1 covariates, e.g. exact sampling time.
Systematic effects

Solution for Fixed Effects

| Effect   | year | timeday | Estimate | Error  | DF  | t Value | Pr > |t|  | Alpha | Lower  | Upper  |
|----------|------|---------|----------|--------|-----|---------|------|---|-------|--------|--------|
| Intercept| 2.3916 | 0.02494 | 2382 | 95.88 | <.0001 | 0.05 | 2.3426 | 2.4405 |
| timeday evening | -2.0137 | 0.02869 | 1802 | -70.19 | <.0001 | 0.05 | -2.0699 | -1.9574 |
| timeday morning | 0 | . . . . | . | . . | . . | . . | . . |
| year 2009 | 0.08465 | 0.03016 | 2421 | 2.81 | 0.0051 | 0.05 | 0.02550 | 0.1438 |
| year 2007 | 0 | . . . . | . | . . | . . | . . | . . |

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>timeday</td>
<td>1</td>
<td>1802</td>
<td>4927.33</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>year</td>
<td>1</td>
<td>2421</td>
<td>7.88</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

Cortisol is measured on log-scale. Backtransformation \( \exp(2.0137) \approx 7.49 \) 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!!!
Explained variation ($R^2$)

We consider only the simplest case, i.e. the two-level model

- we have several variances that can be explained.

**Variation within individuals (residual variation):**

- decreases when we include an important level 1 covariate ($x_1$)
- may also decrease when we include an important level 2 covariate ($x_2$).

**Variation between individuals:**

- decreases when we include an important level 2 covariate ($x_2$)
- *may increase or decrease* when we include an important level 1 covariate ($x_1$)

**Total variance** decreases when including an important covariate.
Hypothetical example I

Covariate $x_1$ varies between individuals, and the variation in individual averages ($\bar{y}$) is mostly due to this variation.

Levels of $y$, for fixed $x$ are quite alike:

- $\omega^2_B$ decreases when $x_1$ is included.
Hypothetical example II

Covariate $x_1$ vary between individuals, but the average outcomes ($\bar{y}$) are almost identical:

Levels of $y$, for fixed $x$ are very different:

- $\omega_B^2$ increases when $x$ is included.
Technical explanation*

A balanced design (same number of observations per cluster):

Explicit solution for the two-level model:

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

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

This is deduced from \( E(MS_B) = n\omega_B^2 + \sigma_W^2 \) and \( E(MS_W) = \sigma_W^2 \).
Outline

General repeated measurements

Random effects ANOVA (the two-level model)

Fixed vs random effects

Multi-level models

Ecological fallacy

Comparing measurement methods
Ecological analyses

The easy way of dealing with repeated measurements:

- Compute summary statistics for each cluster/individual.
- Perform a traditional analysis on the sample of summary statistics rightfully assuming that these are independent.

Summary statistics could be:

- Sample mean or standard deviation.
- AUC (area under the curve).
- Intercept and slope of regression line.

**BUT:** Beware of losing important information.
Ecological vs two-level analysis

Blood pressure and social inequity: 15569 women in 17 regions of Malmø.

Covariates:

- Individual (level 1):
  - low educational achievement ($x$) (less than 9 years of school)
  - age group
- Regional (level 2):
  - rate of people with low educational achievement ($z$) from the 'Skåne Council Statistics Office'
    An aggregated covariate.

Abstract

Study objectives—To study geographical differences in diastolic blood pressure and the influence of the social environment (census percentage of people with low educational achievement) on individual diastolic blood pressure level, after controlling for individual age and educational achievement. To compare the results of multilevel and ecological analyses.

Design—Cross sectional analysis performed by multilevel linear regression modelling, with women at the first level and urban areas at the second level, and by single level ecological regression using areas as the unit of analysis.

Setting—Malmö, Sweden (population 250 000).

Participants—15 569 women aged 45 to 73, residing in 17 urban areas, who took part in the Malmö Diet and Cancer Study (1991–1996).
Ecological analysis

Average blood pressure in region vs rate of people with low educational achievement.

Size of circle indicates size of investigation.

Estimated slope: 4.66 (SE 1.42).

Seems an important explanatory variable?!?
Estimates from two-level model

What is the effect of individual educational achievement ($x_1$) vs regional educational achievement ($x_2$)?

<table>
<thead>
<tr>
<th>Included covariates</th>
<th>Estimate (SE)</th>
<th>Variation</th>
<th>$R^2$ (of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_1$ (individual)</td>
<td>$x_2$ (region)</td>
<td>between regions</td>
</tr>
<tr>
<td>none</td>
<td></td>
<td></td>
<td>0.35</td>
</tr>
<tr>
<td>age</td>
<td></td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td>$x_1$, age</td>
<td>1.15 (0.17)</td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>$x_2$, age</td>
<td>-</td>
<td>4.06 (1.35)</td>
<td>0.12</td>
</tr>
<tr>
<td>$x_1$, $x_2$, age</td>
<td>1.09 (0.17)</td>
<td>2.97 (1.25)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 1 in Merlo et al. (2001)
Ecological analysis vs the two-level model

Region as a random effect could only account for 0.36% of the variation in blood pressures (0.35 of 0.35 + 96.03).

Thus, regional variables such as *rate of low-income* will have very little impact on individual blood pressures!

The ecological analysis ’*sums up*’ the individual and the regional effects, but is *not able to distinguish* between the two.

- It overestimates the level 2 effect.
- It cannot be interpreted as a level 1 effect.
Individual vs regional blood pressure

Figure 2  Actual individual association between individual diastolic blood pressure and census percentage of people with low educational achievement in 17 urban areas of Malmö. The same ecological association as in figure 1 is represented as black circles. This figure permits observation of the variance between the areas (a) (see also fig 1) and between the individual women (b).
Example: suicide and religion

Ecological analysis: Percent of suicides increases with percent of protestants in region.

▶ Are protestants more likely to commit suicide?

Two-level model:

<table>
<thead>
<tr>
<th>level</th>
<th>unit</th>
<th>variation</th>
<th>covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>individuals</td>
<td>within region, $\sigma^2_W$</td>
<td>religion, $x$</td>
</tr>
<tr>
<td>2</td>
<td>regions</td>
<td>between regions, $\omega^2_B$</td>
<td>% protestants, $z$</td>
</tr>
</tbody>
</table>

Finding: Interaction between individual effect ($x$) and region covariate ($z$) . . .
Another example: suicide and religion

More suicides among catholics in regions with many protestants.
Outline

General repeated measurements

Random effects ANOVA (the two-level model)

Fixed vs random effects

Multi-level models

Ecological fallacy

Comparing measurement methods
Comparing measurement devices

Example: Peak expiratory flow rate, l/min:

- 17 subjects, 2 measurement devices,
- two replicates with each method.

<table>
<thead>
<tr>
<th>subject</th>
<th>Wright</th>
<th>mini Wright</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y_{1p1}$</td>
<td>$Y_{1p2}$</td>
</tr>
<tr>
<td>id</td>
<td></td>
<td></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
- Repeatability (variability = measurement error)

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 interchangably?
Simple approaches

For reliability of each method separately we could:
  ▶ make Bland Altman plots of differences vs averages.
  ▶ compute limits of agreement, i.e. the 95% normal range of the differences.

For reproducibility (method comparison) we might:
  ▶ compare the averages in a Bland-Altman plot . . . ?
  ▶ Not good - unless you also do averages in clinic!

For both at the same time:
  ▶ Mixed model for variance between and within methods.
Repeatability

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimated bias</th>
<th>95% limits of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wright</td>
<td>-4.94 (-16.11;6.22)</td>
<td>(-52.33;42.45)</td>
</tr>
<tr>
<td>Mini Wright</td>
<td>2.88 (-11.96;17.73)</td>
<td>(-60.11;65.86)</td>
</tr>
</tbody>
</table>
Two-level models

For each method \((i = 1, 2)\) we have a two-level model

\[
Y_{ijk} = \mu_i + a_{ij} + \varepsilon_{ijk}
\]

- \(\mu_i\) population mean as anticipated by method \(i\).
- \(a_{ij}\) deviation of subject \(j\) from population mean, assumed normally distributed \(N(0, \sigma_i^2)\).
- \(\varepsilon_{ijk}\) deviation for replicate \(k\) (measurement error), assumed normally distributed \(N(0, \omega_i^2)\).
PROC MIXED: Stratified analyses

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

method=mini

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>id</td>
<td>12188</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>396.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>453.91</td>
<td>26.9921</td>
<td>16</td>
<td>16.82</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td></td>
<td>396.69</td>
<td>511.13</td>
<td></td>
</tr>
</tbody>
</table>

method=wright

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>id</td>
<td>13683</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>234.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>447.88</td>
<td>28.4914</td>
<td>16</td>
<td>15.72</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td></td>
<td>387.48</td>
<td>500.29</td>
<td></td>
</tr>
</tbody>
</table>
Joint model for both methods

For methods \((i = 1, 2)\):

\[
Y_{ijk} = \mu_i + a_{ij} + \varepsilon_{ijk}
\]

- \(\varepsilon_{ijk}\) assumed normally distributed \(N(0, \omega_i^2)\) and independent across methods.
- \(a_{ij}\) assumed normally distributed \(N(0, \sigma_i^2)\) and correlated with \(\rho = \text{Cor}(a_{i1}, a_{i2})\).

Anticipated means for the same subject ought to look a lot like each other, so the \(a_{ij}\)'s are likely to be correlated across methods.

- Note that SAS models the covariance parameter \(\sigma_{12} = \text{Cov}(a_{1j}, a_{2j}) = \sigma_1 \cdot \sigma_2 \cdot \rho\).
PROC MIXED: Joint analysis

PROC MIXED DATA=wright;
CLASS method id;
MODEL flow=method / SOLUTION CL;
RANDOM method / TYPE=UN SUBJECT=id;
REPEATED / TYPE=simple GROUP=method SUBJECT=id*method;
RUN;

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Group</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>id</td>
<td></td>
<td>12188</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>id</td>
<td></td>
<td>12542</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>id</td>
<td></td>
<td>13683</td>
</tr>
<tr>
<td>Residual</td>
<td>method*id</td>
<td>method mini</td>
<td>396.44</td>
</tr>
<tr>
<td>Residual</td>
<td>method*id</td>
<td>method wright</td>
<td>234.29</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect      | method | Estimate | StdError | DF  | t Value | Pr > |t| | Alpha | Lower | Upper |
|-------------|--------|----------|----------|-----|---------|-------|---|-------|-------|-------|
| Intercept   |        | 447.88   | 28.4914  | 32  | 15.72   | <.0001| 0.05| 389.85 | 505.92|
| method      | mini   | 6.0294   | 8.0532   | 32  | 0.75    | 0.4595| 0.05| -10.3744 | 22.4332|
| method      | wright |         |         |     |         |       |     |        |       |
Repeatability

Typical differences (approximate 95% normal range) between two measurement with the **same method**:

Wright: \( \hat{\omega}_1^2 = 234.29 \rightarrow \pm 2\sqrt{2\hat{\omega}_1^2} \approx \pm 43.3 \)

Mini: \( \hat{\omega}_2^2 = 396.44 \rightarrow \pm 2\sqrt{2\hat{\omega}_2^2} \approx \pm 56.3 \)

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

\[
F = \frac{396.44}{234.29} = 1.69 \sim F(17, 17) \rightarrow P = 0.14
\]

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

No evidence of **systematic** differences between the two methods.

- Estimated bias +6.0 (-10.4;22.4) for mini vs wright. P=0.46.

Typical differences between the two methods:

\[
\text{var}(Y_{1jk} - Y_{2jk}) = \text{var}(a_{1j} - a_{2j} + \varepsilon_{1jk} - \varepsilon_{2jk}) \\
= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} + \omega_1^2 + \omega_2^2 \\
= 12188 + 13683 - 2 \cdot 12542 + 396.44 + 234.29 \\
= 1417.73
\]

**Limits-of-agreement:** \(6.03 \pm 2\sqrt{1417.7} = (-69.3, 81.3).\)
Not a multi-level 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:

$$Y_{ijk} = \mu_j + a_i + b_{ij} + \varepsilon_{ijk}$$

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

This is assuming the same variance for both methods.
Estimated variance components

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

Covariance Parameter Estimates

<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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Res Log Likelihood</td>
<td>676.0</td>
</tr>
<tr>
<td>AIC (smaller is better)</td>
<td>681.6</td>
</tr>
</tbody>
</table>

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

Between replicate measurements using the same method:

\[ Y_{ijk_1} - Y_{ijk_2} = \varepsilon_{ijk_1} - \varepsilon_{ijk_2} \sim \mathcal{N}(0, 2\sigma^2) \]

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

Between measurements using the different methods:

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

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

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

Solution for Fixed Effects

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

Conclusion: No evidence of systematic differences between the measurement methods.

BUT: Do we really want to assume that variances are equal when the power for testing this is poor?