Recap: linear mixed models and their variance/covariance structure

Rasmus Waagepetersen

October 9, 2024

Model specification

Mixed model = fixed effects + random effects

With `lmer`:

```
y~‘fixed formula’+(‘rand formula_1’|Group_1)+ ...
                             +(‘rand. formula_n’|Group_n)
```

Fixed effects just like ordinary multiple regression (`lm()`)

Important feature of linear mixed models: using simple building blocks (independent random effects) we can obtain complex and more realistic models for the covariance structure of our observations.

Very wide range of models possible but one should carefully consider what makes sense for the particular data and research question considered.

Easy to specify way too complex models !

Example from last time

We may get "strange" output:

```
> ort7=lmer(distance~age*Sex+(age|Subject))
> summary(ort7)
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 5.77441  2.4030
          age         0.03245  0.1801   -0.67
 Residual             1.71661  1.3102
Number of obs: 108, groups:  Subject, 27
```

This is model with correlated (estimate -0.67 for ρ) subject specific intercept and slope.

Child with big intercept has small slope and vice versa

How do results comply with other analyses which say that total variance $\mathbb{V}\mathrm{ar}\, Y$ a bit more than 5 ?

Why care about theoretical calculations of variance ?

If we understand the basics of variance and covariance calculations we can understand previous output !

Random coefficient models (exercise for last time):

$$Y = \alpha + a + \beta x + \epsilon \qquad \mathbb{V}\mathrm{ar}\, Y = \tau_a^2 + \sigma^2$$
$$Y = \alpha + [\beta + b]x + \epsilon \qquad \mathbb{V}\mathrm{ar}\, Y = x^2\tau_b^2 + \sigma^2$$
$$Y = \alpha + a + [\beta + b]x + \epsilon \qquad \mathbb{V}\mathrm{ar}\, Y = \tau_a^2 + x^2\tau_b^2 + 2x\mathbb{C}\mathrm{ov}(a, b) + \sigma^2$$

NB: for the last two models, $\mathbb{V}\mathrm{ar}\, Y$ is a ‘smiling’ second order polynomial in the covariate x !

Variance is sum of covariances between random terms:

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\, Y = \mathbb{V}\mathrm{ar}(\alpha + a + \beta x + bx + \epsilon) = \\
&\mathbb{C}\mathrm{ov}(\alpha + a + \beta x + bx + \epsilon, \alpha + a + \beta x + bx + \epsilon) = \\
&\mathbb{C}\mathrm{ov}(a, a) + \mathbb{C}\mathrm{ov}(a, bx) + \mathbb{C}\mathrm{ov}(bx, bx) + \mathbb{C}\mathrm{ov}(bx, a) + \mathbb{C}\mathrm{ov}(\epsilon, \epsilon) = \\
&\mathbb{V}\mathrm{ar}(a) + \mathbb{V}\mathrm{ar}(bx) + \mathbb{V}\mathrm{ar}\epsilon + 2x\mathbb{C}\mathrm{ov}(a, b) = \\
&\tau_a^2 + x^2\tau_b^2 + \sigma^2 + 2x\rho\tau_a\tau_b
\end{aligned}
$$

NB: a and b are assumed to be independent of ϵ so e.g. $\mathbb{C}\mathrm{ov}(a, \epsilon) = 0$

NB: we may or may not assume a and b to be independent.

Correlation and covariance:

$$
\rho = \mathbb{C}\mathrm{orr}(a, b) = \frac{\mathbb{C}\mathrm{ov}(a, b)}{\sqrt{\mathbb{V}\mathrm{ar}\, a}\sqrt{\mathbb{V}\mathrm{ar}\, b}} \Rightarrow \mathbb{C}\mathrm{ov}(a, b) = \rho\tau_a\tau_b
$$

We have from previous slide:

$$\mathbb{V}\mathrm{ar}\, Y = \tau_a^2 + x^2\tau_b^2 + 2x\mathbb{C}\mathrm{ov}(a,b) + \sigma^2 \quad \mathbb{C}\mathrm{ov}(a,b) = \rho\tau_a\tau_b$$

Using output:

$$\mathbb{C}\mathrm{ov}(a,b) = 2.40 * 0.18 * (-0.67) = -0.28$$

Age 8:

$$\mathbb{V}\mathrm{ar}\, Y = 5.77 + 8^2 * 0.032 + 2 * 8 * (-0.28) + 1.72 = 5.06$$

Age 10:

$$\mathbb{V}\mathrm{ar}\, Y = 5.77 + 10^2 * 0.032 + 2 * 10 * (-0.28) + 1.72 = 5.09$$

Age 12:

$$\mathbb{V}\mathrm{ar}\, Y = 5.77 + 12^2 * 0.032 + 2 * 12 * (-0.28) + 1.72 = 5.38$$

Age 14:

$$\mathbb{V}\mathrm{ar}\, Y = 5.77 + 14^2 * 0.032 + 2 * 14 * (-0.28) + 1.72 = 5.92$$

Variances increase with age but in agreement with other analyses (multiple regression, linear mixed with random intercepts) in terms of total variance.

Sugar beets example

Outcome: sugar percentage

Two treatments: harvest time and sowing time.

Experimental design: 6 plots organized in 3 blocks. 5 split-plots within each plot. In total 30 observations.[1]

```
Harvesting dates:
  1: 2/10, 2: 21/10

Plot allocation:
             |  Block 1        |  Block 2        |  Block 3        | Time
             +-----------------|-----------------|-----------------+
Split-plots  | h1 h1 h1 h1 h1  | h2 h2 h2 h2 h2  | h1 h1 h1 h1 h1  | Harvesting
 1-15        | s3 s4 s5 s2 s1  | s3 s2 s4 s5 s1  | s5 s2 s3 s4 s1  | Sowing
             ------------------|-----------------|-----------------|
Split-plots  | h2 h2 h2 h2 h2  | h1 h1 h1 h1 h1  | h2 h2 h2 h2 h2  | Harvesting
16-30        | s2 s1 s5 s4 s3  | s4 s1 s3 s2 s5  | s1 s4 s3 s2 s5  | Sowing
             +-----------------|-----------------|-----------------+
```

[1] figure reproduced from Halehoh and Højsgaard (2014)

Linear mixed model ?

We can use indices $b = 1, 2, 3$ for block, $h = 1, 2$ for harvest time and $s = 1, \dots, 5$ for sowing time.

Which linear mixed effects model should we use ?

Which fixed effects - which random effects ?

What about nurse examples (exercise 4 from slides 1) ?

Why you should not add standard deviations

Consider one-way anova with random effects.

Total variance is

$$\mathbb{V}\mathrm{ar}\, Y_{ij} = \tau^2 + \sigma^2$$

Total standard deviation is

$$\sqrt{\mathbb{V}\mathrm{ar}\, Y_{ij}} = \sqrt{\tau^2 + \sigma^2} \neq \tau + \sigma$$

For example (Pythagoras)

$$5^2 = 3^2 + 4^2 \text{ but } 5 \neq 3 + 4$$

Model with subject specific intercepts

```
> ortss=lm(distance~-1+Subject+age+age:factor(Sex)+factor(Sex))
> summary(ortss)

Coefficients: (1 not defined because of singularities)
Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
SubjectM16            14.3719     1.0988  13.080  < 2e-16 ***
SubjectM05            14.3719     1.0988  13.080  < 2e-16 ***
SubjectM02            14.7469     1.0988  13.421  < 2e-16 ***
SubjectM11            14.9969     1.0988  13.649  < 2e-16 ***
SubjectM07            15.1219     1.0988  13.763  < 2e-16 ***
SubjectM08            15.2469     1.0988  13.876  < 2e-16 ***
SubjectM03            15.6219     1.0988  14.218  < 2e-16 ***
SubjectM12            15.6219     1.0988  14.218  < 2e-16 ***
...
SubjectF01            16.1000     1.2400  12.984  < 2e-16 ***
SubjectF05            17.3500     1.2400  13.992  < 2e-16 ***
SubjectF07            17.7250     1.2400  14.294  < 2e-16 ***
SubjectF02            17.7250     1.2400  14.294  < 2e-16 ***
SubjectF08            18.1000     1.2400  14.597  < 2e-16 ***
SubjectF03            18.4750     1.2400  14.899  < 2e-16 ***
SubjectF04            19.6000     1.2400  15.806  < 2e-16 ***
SubjectF11            21.1000     1.2400  17.016  < 2e-16 ***
age                    0.7844     0.0775  10.121 6.44e-16 ***
factor(Sex)Female          NA         NA      NA       NA
age:factor(Sex)Female -0.3048     0.1214  -2.511   0.0141 *
```

NB: omitted common intercept (-1 in model formula)

For each subject an estimate of deviation between the subject's intercept and the first subject's intercept.

In total 27 (!) subject specific estimates.

Each estimate pretty poor (only 4 observations for each subject).

Can not estimate female effect !

Model with subject specific effects may be more correct but is it useful ?

Overparametrization for orthodont data

Model with subject specific intercepts:

$$Y_{ij} = \mu + \alpha_i + \delta_{\text{sex}(i)} + \beta \text{age}_{ij} + \beta_{\text{sex}(i)} \text{age}_{ij} + \epsilon_{ij}$$

Why can't we estimate $\delta_{\text{sex}(i)}$?

sex(i) is 1 if individual i is a girl and 0 otherwise.

Note for any constant c we have

$$\alpha_i + \delta_{\text{sex}(i)} = (\alpha_i + c) + (\delta_{\text{sex}(i)} - c)$$

In words: if we increase all girl intercepts by c and decrease the sex effect by c the expected value of Y_{ij} is unchanged.

Thus we can not identify a unique best fitting value of δ_1

We do not have this problem if α_i is substituted by random effect U_i which is not used to model the expected value of an observation.