Modeling with random effects

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 18, 2024

Course topics

- random effects
- linear mixed models
- statistical inference for linear mixed models (including analysis of variance)
- prediction of random effects
- Implementation in R and SPSS

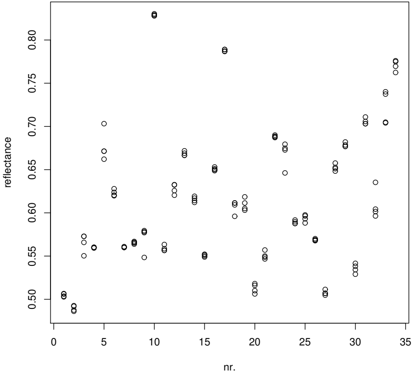
Outline - first session

- examples of data sets
- random effects models - motivation and interpretation

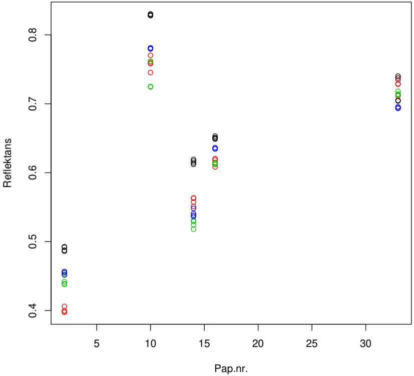
Next session : details on implementation in R and SPSS

Reflectance (colour) measurements for samples of cardboard (egg trays) (project at Department of Biotechnology, Chemistry and Environmental Engineering)

Four replications at same position on each cardboard



For five cardboards: four replications at four positions at each cardboard

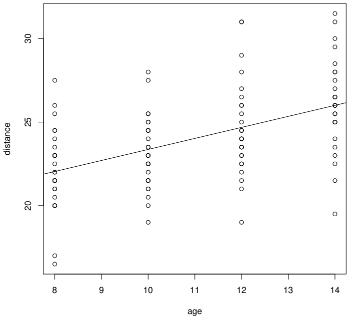


Colour variation between/within cardboards ?

Orthodontic growth curves (repeated measurements/longitudinal data)

Distance (related to jaw size) between pituitary gland and the pterygomaxillary fissure (two distinct points on human skull) for children of age 8-14

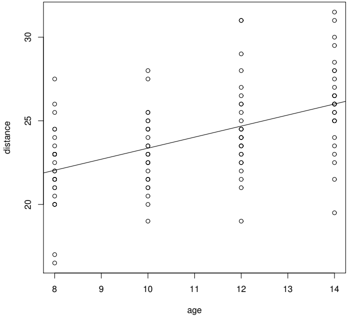
Distance versus age:



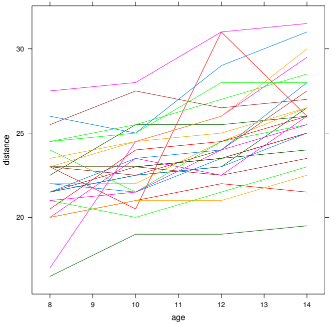
Orthodontic growth curves (repeated measurements/longitudinal data)

Distance (related to jaw size) between pituitary gland and the pterygomaxillary fissure (two distinct points on human skull) for children of age 8-14

Distance versus age:

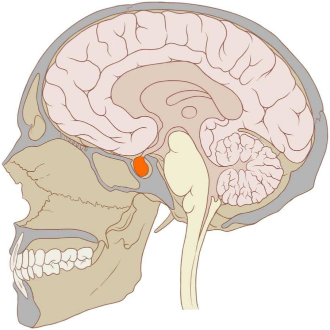


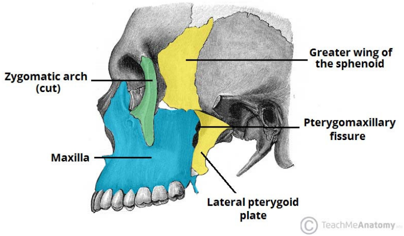
Distance versus age grouped according to child



Different intercepts for different children !

(anatomy of the scull)





Pituitary gland is orange object.

Whole grain (WG) vs. refined grain (RG)

Outcome: LDL cholesterol in blood

Subjects randomly allocated to two treatment groups. Three measurements for each subject:

Group 1:	baseline	WG	RG
Group 2:	baseline	RG	WG

Note: possible cross over effect (treatment effect WG-RG may depend on order of treament (WG first or last)

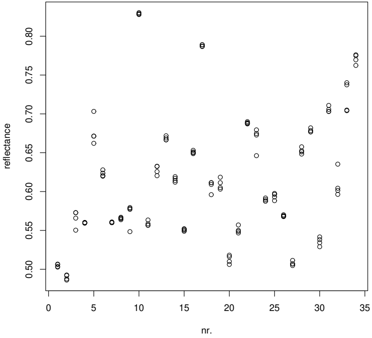
Outcome may vary a lot between subjects with same treatment.

Recall: basic aim for statistical analysis of a sample/dataset is to extract information that can be generalized to the population that was sampled.

This perspective in mind when deciding on models for the datasets considered.

Model for reflectances: one-way anova

Four replications on each cardboard



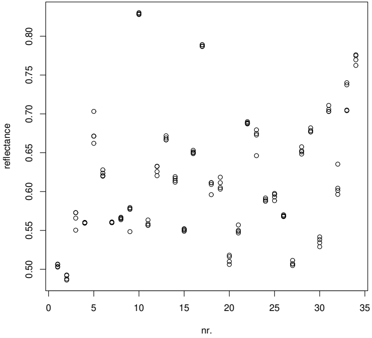
Models:

$$Y_{ij} = \mu + \epsilon_{ij} \quad i = 1, \ldots, k \;\; j = 1, \ldots, m$$

($k = 34$, $m = 4$) where μ expectation and ϵ_{ij} random independent noise

Model for reflectances: one-way anova

Four replications on each cardboard



Models:

$$Y_{ij} = \mu + \epsilon_{ij} \quad i = 1, \ldots, k \;\; j = 1, \ldots, m$$

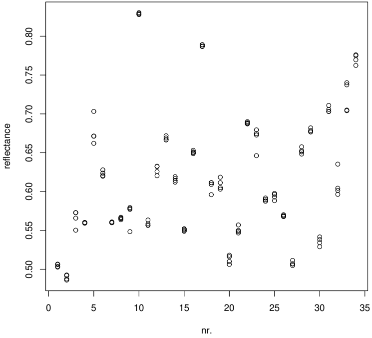
($k = 34,\ m = 4$) where μ expectation and ϵ_{ij} random independent noise or

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where α_i are fixed unknown parameters

Model for reflectances: one-way anova

Four replications on each cardboard



Models:

$$Y_{ij} = \mu + \epsilon_{ij} \quad i = 1, \ldots, k \;\; j = 1, \ldots, m$$

$(k = 34,\ m = 4)$ where μ expectation and ϵ_{ij} random independent noise or

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where α_i are fixed unknown parameters or

$$Y_{ij} = \mu + U_i + \epsilon_{ij}$$

where U_i are zero-mean random variables independent of each other and of ϵ_{ij}

Which is most relevant ?

One role of random effects: parsimonious and population relevant models

With fixed effects α_i: many parameters (μ, σ^2, $\alpha_1, \dots, \alpha_{34}$). Parameters $\alpha_1, \dots, \alpha_{34}$ not interesting as they just represent intercepts for specific card boards which are individually not of interest.

With random effects: just three parameters (μ, $\sigma^2 = \mathbb{V}\mathrm{ar}\epsilon_{ij}$ and $\tau^2 = \mathbb{V}\mathrm{ar}U_i$).

Hence parsimonious model. Variance parameters interesting for several reasons.

Second role of random effects: quantify sources of variation

Quantify sources of variation (e.g. quality control): is pulp for paper production too heterogeneous ?

With random effects model

$$Y_{ij} = \mu + U_i + \epsilon_{ij} \tag{1}$$

we have decomposition of variance:

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

Hence we can quantify variation between (τ^2) cardboard pieces and within (σ^2) cardboard.

Ratio $\gamma = \tau^2/\sigma^2$ is 'signal to noise'.

Proportion of variance

$$\frac{\tau^2}{\sigma^2 + \tau^2} = \frac{\gamma}{\gamma + 1}$$

is called *intra-class correlation*.

High proportion of between cardboard variance leads to high correlation (next slide).

Third role: modeling of covariance and correlation

Covariances:

$$\mathbb{C}\mathrm{ov}[Y_{ij}, Y_{lk}] = \begin{cases} 0 & i \neq l \\ \mathbb{V}\mathrm{ar} U_i = \tau^2 & i = l, j \neq k \\ \mathbb{V}\mathrm{ar} U_i + \mathbb{V}\mathrm{ar} \epsilon_{ij} = \tau^2 + \sigma^2 & i = l, j = k \end{cases} \tag{2}$$

Correlations:

$$\mathbb{C}\mathrm{orr}[Y_{ij}, Y_{lk}] = \begin{cases} 0 & i \neq l \\ \tau^2/(\sigma^2 + \tau^2) & i = l, j \neq k \\ 1 & i = l, j = k \end{cases} \tag{3}$$

That is, observations for same cardboard are correlated !

Correct modeling of correlation is important for correct evaluation of uncertainty.

Fourth role: correct evalution of uncertainty

Suppose we wish to estimate $\mu = \mathbb{E} Y_{ij}$. Due to correlation, observations on same cardboard to some extent redundant.

Estimate is empirical average $\hat{\mu} = \bar{Y}_{\cdot\cdot}$. Evaluation of $\mathbb{V}\mathrm{ar}\,\bar{Y}_{\cdot\cdot}$:

Model erroneously ignoring variation between cardboards

$$Y_{ij} = \mu + \epsilon_{ij}$$

$$\mathbb{V}\mathrm{ar}\epsilon_{ij} = \sigma^2_{\text{total}} \left[= \sigma^2 + \tau^2\right]$$

Naive variance expression is

$$\mathbb{V}\mathrm{ar}\,\bar{Y}_{\cdot\cdot} = \frac{\sigma^2_{\text{total}}}{n} \left[= \frac{\sigma^2 + \tau^2}{mk}\right]$$

Correct model with random cardboard effects

$$Y_{ij} = \mu + U_i + \epsilon_{ij},$$

$$\mathbb{V}\mathrm{ar}\,U_i = \tau^2, \quad \mathbb{V}\mathrm{ar}\epsilon_{ij} = \sigma^2$$

Correct variance expression is

$$\mathbb{V}\mathrm{ar}\,\bar{Y}_{\cdot\cdot} = \frac{\tau^2}{k} + \frac{\sigma^2}{mk} \qquad (4)$$

With first model, variance is underestimated !

For $\mathbb{V}\mathrm{ar}\,\bar{Y}_{\cdot\cdot} \to 0$ is it enough that $mk \to \infty$?

Whole grain (WG) vs. refined grain (RG) - model

For ith subject three measurements Y_{it}, $t = 1, 2, 3$

Standard approach: regression using baseline Y_{1t} as covariate (to correct for person-specific effects):

$$Y_{it} = \mu_{it} + \alpha Y_{i1} + \epsilon_{it}, \quad t = 2, 3$$

μ_{it}: mean depends on Group (1, 2) and Treatment (WG, RG)

Problem: we need to skip all observations for i if baseline is missing !

Alternative: mixed model with subject specific random effect

$$Y_{it} = \mu_{it} + U_i + \epsilon_{it}, \quad t = 1, 2, 3$$

Classical balanced one-way ANOVA (analysis of variance)

Decomposition of empirical variance/sums of squares ($i = 1, \ldots, k$, $j = 1, \ldots, m$):

$$SST = \sum_{ij} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{ij} (Y_{ij} - \bar{Y}_{i\cdot})^2 + m \sum_{i} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = SSE + SSB$$

Expected sums of squares:

$$\mathbb{E}SSE = k(m-1)\sigma^2$$

$$\mathbb{E}SSB = m(k-1)\tau^2 + (k-1)\sigma^2$$

Moment-based estimates:

$$\hat{\sigma}^2 = \frac{SSE}{k(m-1)} \quad \hat{\tau}^2 = \frac{SSB/(k-1) - \hat{\sigma}^2}{m}$$

More complicated formulae in the unbalanced case.

Hypothesis tests

Fixed effects: H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

$$F = \frac{SSB/(k-1)}{SSE/(k(m-1))}$$

Random effects: H_0: $\tau^2 = 0$ Same test-statistic

$$F = \frac{SSB/(k-1)}{SSE/(k(m-1))}$$

Idea: if τ^2=0 then $\mathbb{E}SSB/(k-1) = \mathbb{E}SSE/(k(m-1)) = \sigma^2$.
Hence under H_0, F should be close to 1.

If $\tau^2 > 0$ then
$\mathbb{E}SSB/(k-1) = m\tau^2 + \sigma^2 > \mathbb{E}SSE/(k(m-1)) = \sigma^2$. Thus big values of F critical for H_0.

Classical implementation in R

For cardboard/reflectance data, $k = 34$ and $m = 4$. `anova()` procedure produces table of sums of squares.

```
> anova(lm(Reflektans~factor(Pap.nr.)))
Analysis of Variance Table

Response: Reflektans
               Df  Sum Sq Mean Sq F value
factor(Pap.nr) 33  0.9009 0.0273  470.7    #SSB
Residuals      102 0.0059 0.00006          #SSE
---
```

Hence $\hat{\sigma}^2 = 0.00006$, $\hat{\tau}^2 = (0.0273 - 0.00006)/4 = 0.00681$.

Biggest part of variation is between cardboard.

Orthodontic data: classical multiple linear regression in R

```
#fit model with sex specific intercepts and slopes
> ort1=lm(distance~age+age:factor(Sex)+factor(Sex))
> summary(ort1)
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            16.3406     1.4162  11.538  < 2e-16 ***
age                     0.7844     0.1262   6.217 1.07e-08 ***
factor(Sex)Female       1.0321     2.2188   0.465    0.643
age:factor(Sex)Female  -0.3048     0.1977  -1.542    0.126
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.257 on 104 degrees of freedom
Multiple R-squared: 0.4227,Adjusted R-squared: 0.4061
F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```

Sex and age:Sex not significant !

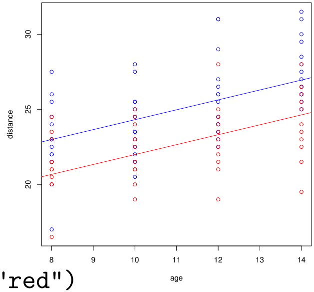
Multiple linear regression continued - without interaction

```
> ort2=lm(distance~age+factor(Sex))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        17.70671    1.11221  15.920  < 2e-16 ***
age                 0.66019    0.09776   6.753 8.25e-10 ***
factor(Sex)Female -2.32102    0.44489  -5.217 9.20e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 2.272 on 105 degrees of freedom
Multiple R-squared: 0.4095,Adjusted R-squared: 0.3983
F-statistic: 36.41 on 2 and 105 DF,  p-value: 9.726e-13
```

both age and sex significant

Multiple linear regression in R III

```
#plot data and two regression lines
col=rep("blue",length(Sex))
col[Sex=="Female"]="red"
plot(distance~age,col=col)
abline(parm[1:2],col="blue")
abline(c(parm[1]+parm[3],parm[2]),col="red")
```



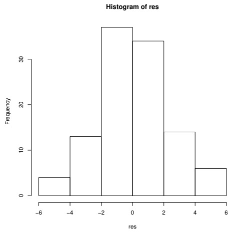
Multiple linear regression in R IV

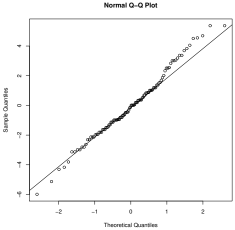
```
res=residuals(ort2)
```

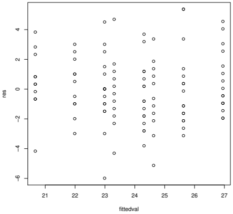
```
hist(res)
```

```
qqnorm(res)
qqline(res)
```

```
fittedval=fitted(ort
plot(res~fittedval)
```

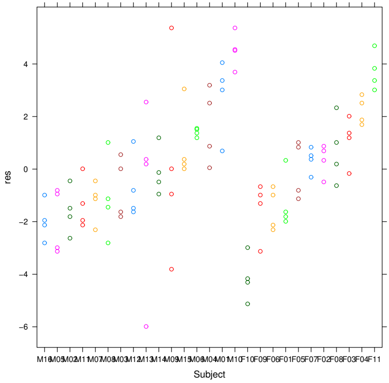






Multiple linear regression in R V

```
> library(lattice)
> xyplot(res~Subject,groups=Subject)
```



Oups - residuals not independent and identically distributed ! Hence computed F-tests not valid.

Problem: subject specific intercepts (and possibly subject specific slopes too)

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.

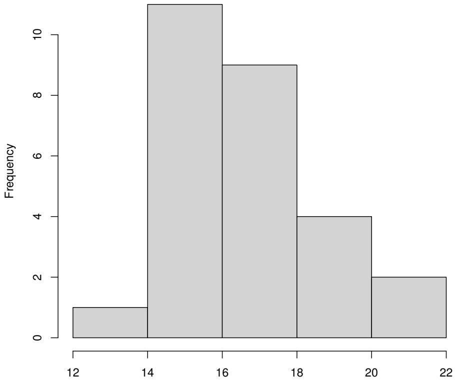
In total 27 (!) subject specific estimates.

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 ?

Distribution of estimates of subject specific effects



Normal distribution for subject specific intercepts ?

Mixed model for growth data

$$Y_{ij} = \alpha + \delta_{\text{sex}(i)} + \beta x_{ij} + a_i + b_i x_{ij} + \epsilon_{ij}, \quad i\text{: child}, j\text{: time}$$

Models for coefficients:

- If interest lies in mean intercept and slope (α, β) and sex difference δ_s but not individual subjects then wasteful to include subject specific fixed effects a_i and b_i (want parsimonious models).
- Using random effects a_i and b_i with variances τ_a^2 and τ_b^2 allows quantification of population heterogeneity. And only unknown parameters α, β, δ_s, τ_a^2, τ_b^2 and σ^2 (do not need to estimate a_i and b_i)

Back to first role of random effects: parsimonious and meaningful modeling of heterogeneous data.

Mixed model: both systematic and random effects.

Marginal and conditional means of observations

Suppose $a_i \sim N(0, \tau_a^2)$ and $b_i \sim N(0, \tau_b^2)$

Unconditional (marginal) mean of observation:

$$\mathbb{E}[Y_{ij}] = \alpha + \delta_{\text{sex}(i)} + \beta \text{age}_{ij}$$

- i.e. one regression line for each sex (population mean of subject specific lines).

Conditional on a_i and b_i:

$$\mathbb{E}[Y_{ij} | a_i, b_i] = [\alpha + a_i] + \delta_{\text{sex}(i)} + [\beta + b_i] \text{age}_{ij}$$

i.e. subject specific lines vary randomly around population mean.

Mixed model analysis of orthodont data

```
> ort4=lmer(distance~age+Sex+(1|Subject))
> summary(ort4)
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 3.2668   1.8074
 Residual             2.0495   1.4316
Number of obs: 108, groups: Subject, 27

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 17.70671    0.83392 99.35237  21.233  < 2e-16 *
age          0.66019    0.06161 80.00000  10.716  < 2e-16 *
SexFemale   -2.32102    0.76142 25.00000  -3.048  0.00538 *
```

Both age and Sex significant. Estimates coincide with those for linear regression but larger standard error for Sex.

Comparison of variances

Between subject variance: 3.27, Noise variance: 2.05.

Total variance: 3.27+2.05=5.32

Similar to estimated residual variance for multiple linear regression model: $5.26 = 2.272^2$.

Looking at interaction in mixed model framework

```
Formula: distance ~ age * Sex + (1 | Subject)

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 3.299    1.816
 Residual             1.922    1.386
Number of obs: 108, groups:  Subject, 27

Fixed effects:
              Estimate Std. Error       df t value Pr(>|t|)
(Intercept)    16.3406     0.9813 103.9864  16.652  < 2e-16 ***
age             0.7844     0.0775  79.0000  10.121 6.44e-16 ***
SexFemale       1.0321     1.5374 103.9864   0.671   0.5035
age:SexFemale  -0.3048     0.1214  79.0000  -2.511   0.0141 *
```

Now interaction significant !

What is interpretation of interaction ? Does it make sense ?

Note: corresponding model without random effects has much inflated residual variance $5.09 = 2.257^2$ vs. 1.922 for mixed model.

Interaction 'drowns' in large random noise.

Summary - role of random effects

Models with random effects (mixed models) are useful for:

- quantifying different sources of variation
- appropriate modeling of variance structure and correlation
- correct evalution of uncertainty of parameter estimates
- estimation of population variation instead of subject specific characteristics
- more parsimonious models (one variance parameter vs. many subject specific fixed effects parameters)

Exercises

For exercises 1 and 3 recall:

$$\mathbb{C}\mathrm{ov}(X_1 + X_2 + \cdots + X_n, Y_1 + Y_2 + \cdots + Y_m)$$
$$= \mathbb{C}\mathrm{ov}(X_1, Y_1) + \mathbb{C}\mathrm{ov}(X_1, Y_2) + \cdots + \mathbb{C}\mathrm{ov}(X_n, Y_m)$$

Also recall if either X_i or Y_j is non-random or X_i and X_j independent then $\mathbb{C}\mathrm{ov}(X_i, Y_j) = 0$.

1. Show results regarding covariances and correlations in equations (2) and (3) for the Y_{ij} in one-way ANOVA (i.e. the model in equation (1)).
2. Analyze the pulp data (brightness of paper pulp in groups given by different operators; from the faraway package) using a one-way anova with random operator effects. Estimate variance components and the intra-class correlation (you may also use output on next slide).

One-way anova for pulp data (4 operators, 5 observations for each operator):

```
> anova(lm(bright~operator,data=pulp))
Analysis of Variance Table

Response: bright
          Df Sum Sq Mean Sq F value  Pr(>F)
operator   3   1.34 0.44667  4.2039 0.02261 * #SSB
Residuals 16   1.70 0.10625                   #SSE
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

More exercises

3. In this exercise α and β are non-random parameters. Also x_{ij} is considered non-random (the linear regressions are models for Y_{ij} *conditional* on x_{ij}).
   3.1 Compute variance of observations from the linear model with random intercepts:

   $$Y_{ij} = \alpha + a_i + \beta x_{ij} + \epsilon_{ij}$$

   where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $a_i \sim N(0, \tau_a^2)$ and the ϵ_{ij} and a_i are independent.
   3.2 Consider the model fitted on slide 'Mixed model analysis of orthodont data'. What is the proportion of variance due to the error (residual) term ?
   3.3 Compute variances, covariances and correlations of observations from the linear model with random slopes:

   $$Y_{ij} = \alpha + \beta x_{ij} + b_i x_{ij} + \epsilon_{ij}$$

   where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $b_i \sim N(0, \tau_b^2)$ and the ϵ_{ij} and b_i are independent.

3. 3.4 Consider following output. What is the proportion of variance for an observation Y_{ij} explained by the random slopes for different values 8, 10, 12, and 14 of age ?

```
> ort5=lmer(distance~age+Sex+(-1+age|Subject))
> summary(ort5)

Random effects:
 Groups   Name Variance Std.Dev.
 Subject  age  0.026374 0.1624
 Residual      2.080401 1.4424
Number of obs: 108, groups: Subject, 27

Fixed effects:
            Estimate Std. Error t value
(Intercept) 17.43042    0.75066  23.220
age          0.66019    0.06949   9.500
SexFemale   -1.64286    0.68579  -2.396
```

4. Consider the following examples. Is there scope for using random effects - and if so, how ?
   4.1 In an agricultural experiment 2 different varieties of barley and two types A and B of fertilizer are tried out on 10 fields. Each variety is applied to 5 fields where the allocation of varieties to fields is random. Each field is further split into two plots where one part receives fertilizer A and the other fertilizer B. The dependent variable is barley yield within plots.
   4.2 10 nurses treat 40 patients where 20 patients receive treatment A and 20 receive treatment B (both against high blood pressure). Each nurse takes care of four patients where two gets treatment A and two gets treatment B. Dependent variable is blood pressure measured once a week over 5 weeks.
   4.3 The experiment in previous question is changed so that only 2 nurses are involved. One nurse treats 20 patients with A and one nurse treats 20 patients with B. Again blood pressure is measured 5 times for each patient (extra question: is this a good design ?)
   4.4 What is the implication for estimation of variances if there is just one blood pressure measurement for each patient ? Do you prefer to include 10 or 2 nurses ?

5. compute $\mathbb{V}\mathrm{ar}\,\bar{Y}_{..}$ for one way ANOVA (equation (4)).