Exercise 2 Comparison of two infant formulas

This exercise concerns data of a randomized double blind clinical trial comparing two infant formulas (milk). The outcome variable considered in this exercise is:

Y = daily intake of the infant formula (x100 ml)

Mothers visited a clinic at 4, 8, 13 and 17 weeks of age of their infant, but occasionally visits were missing. The mothers had to register the daily intake of the formula in a diary during 7 days before each visit, but in practice the number of days of a diary varied between 3 and 9 days. The infants were randomized at the time that the mother decided to stop full breast feeding and to start formula feeding. The age of randomization therefore varied between infants and the data set contains only the visits after the randomization. The time variable in the analysis is the time (in days) since randomization. On this time scale the outcome measurements are very unbalanced.

To illustrate the structure of the data, the data of infant 2 are given in the next table. Time is the number of days since randomization, and group denotes the formula (0 = control; 1 = experimental). The first visit for infant 2 (subject 2) is missing since that took place before the randomization. Notice that the number of diary days varies among visits, e.g. 7, 6, and 5.

id	time	visit	group	y.100
2	20	2	1	8.77
2	21	2	1	8.77
2	22	2	1	8.80
2	23	2	1	8.80
2	24	2	1	8.82
2	25	2	1	8.85
2	26	2	1	8.82
2	70	3	1	9.89
2	71	3	1	9.85
2	72	3	1	9.95
2	73	3	1	9.90
2	74	3	1	9.96
2	75	3	1	9.93
2	90	4	1	10.73
2	91	4	1	10.70
2	92	4	1	10.72
2	93	4	1	10.72
2	94	4	1	10.72

For the first two of three analyses we calculate the average of Y (`meany`) and the corresponding average day per visit (`meantime`). For instance, in this aggregated dataset the data of subject 2 reduce to:

id	meantime	visit	meany
2	23.0	2	8.81
2	72.5	3	9.91
2	92.0	4	10.72

The first two mixed models (models 1 and 2) are fitted on this aggregated data set. Answer the following questions for **model 1** with the output for model 1 at the end of this exercise.

a. **[1.5]** Introduce your own notation and present the mathematical formula for the mixed model used.

b. **[1.5]** Specify the model assumptions.

Solution:

$y_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 group_i + \beta_3 time_{ij} group_i + b_i + \varepsilon_{ij}$,

$i = 1, \ldots, n$ denotes the mothers,
$j = 1, \ldots, 4$ denotes the visits
y_{ij} is the aggregated daily intake for mother i at visit j
$\varepsilon_{ij} \sim N(0, \sigma^2)$ *and* $b_i \sim N(0, \sigma_b^2)$
ε_{ij} is independent of b_i
σ^2 is the measurement error variance
σ_b^2 is the random effects variance

Note that explanation of the subscripts is also expected.

c. **[1.0]** What is the size of the estimated difference between the two groups in mean aggregated daily intake at 100 days after randomization?

Solution:

Based on the model above, it is: $\beta_2 + 100\beta_3$*=1.42+100*0.40=41.42.*

Note that it is not enough to give the formula; you are also expected to do the computation.

d. **[3.0]** What is the estimated variance at a mean time of 15 days after randomization? Is the variance in this model different for different (mean) times?
Solution:

Estimated variance = 5.953944^2+ 2.232729^2 = 40.43453, independent of the time.

Note that a common mistake is the use of the standard deviation instead of the variances and not both questions are answered.

e. **[2.0]** What is the estimated correlation between measurements at (mean) times 15 and 60 days done at the same infant? Does this correlation depend on the chosen pair of times?

Solution:

Correlation = 5.953944^2/(5.953944^2+ 2.232729^2) =0.8767123, independent of the times.

Note that a common mistake is the use of the standard deviation instead of the variances and not both questions are answered.

f. **[2.0]** Model 1 above has been fitted with the R package 'nlme' and the code used is:

```
library(nlme)
model. aggr <- lme(meany ~ meantime+ group + meantime:group,
                   random = ~ 1|id,
                   data = data.aggr2)
```

If we used instead the following code:

```
model.aggr2 <- gls(meany ~ meantime+ group + meantime:group,
                   correlation = corCompSymm(form = ~ 1 | id),
                   data = data.aggr2)
```

would that lead to another model or would it be identical? Explain and motivate your answer.

Solution:

Compound Symmetry also specifies equal variances and equal correlations independent of time, so the models would be identical.

g. **[6.0]** What is the interpretation of the regression coefficient of **meantime?** Under the "Fixed effects" part of the output and under the "anova(model.aggr)" part of the output a test for the coefficient of **meantime** is given. Are they the same? What is (are) the corresponding null hypothesis (hypotheses)?

Solution:

*The coefficient 1.084706 estimates the increase per day in the mean formula intake **in group 0.** The corresponding P-value is for the test of H_0: increase in mean intake per day in group 0 is equal to 0. The null hypothesis tested under "anova(model.aggr)" part of the output is for H_0: increase in mean intake per day **averaged over group 0 and 1** is equal to 0. Notice that the squared t-statistic is much different from the F-statistic, but the P-values lead to the same inference.*

In **model 2** a second random effect is added. Look at the output of model 2 at the end of this exercise and answer the following questions.

h. **[4.0]** Test the null hypothesis that the variance of the added random effect is zero. Give the test statistic and corresponding degrees of freedom and the P-value (for the P-value it is enough to indicate whether it is nonsignificant, just significant or very significant).

Solution:

REML LR test is:

```
> anova(model.1, model.2)
         Model df      AIC      BIC    logLik   Test     L.Ratio p-value
model.1     1  6 661.2938 676.6799 -324.6469
model.2     2  8 665.2910 685.8058 -324.6455 1 vs 2 0.002767856  0.9986
```

with df = 2:1, not significant.

i. **[8.0]** What is the estimated covariance between measurements at mean times 15 and 20 days done at the same infant?

Solution:

$$cov(b_0 + 15b_1 + \varepsilon, b_0 + 20b_1 + \varepsilon) = var(b_0) + 35 \cdot \text{covar}(b_0, b_1) + 300\, var(b_1) =$$

```
6.394928962^2+35*(-0.999*6.394928962*0.005495021)+300*0.005495021^2= 39.6755
```

j. **[2.0]** Give the design matrix for the random effects terms for a mother who has provided aggregated daily intake data only at 0, 100 and 200 days after randomization.

Solution:

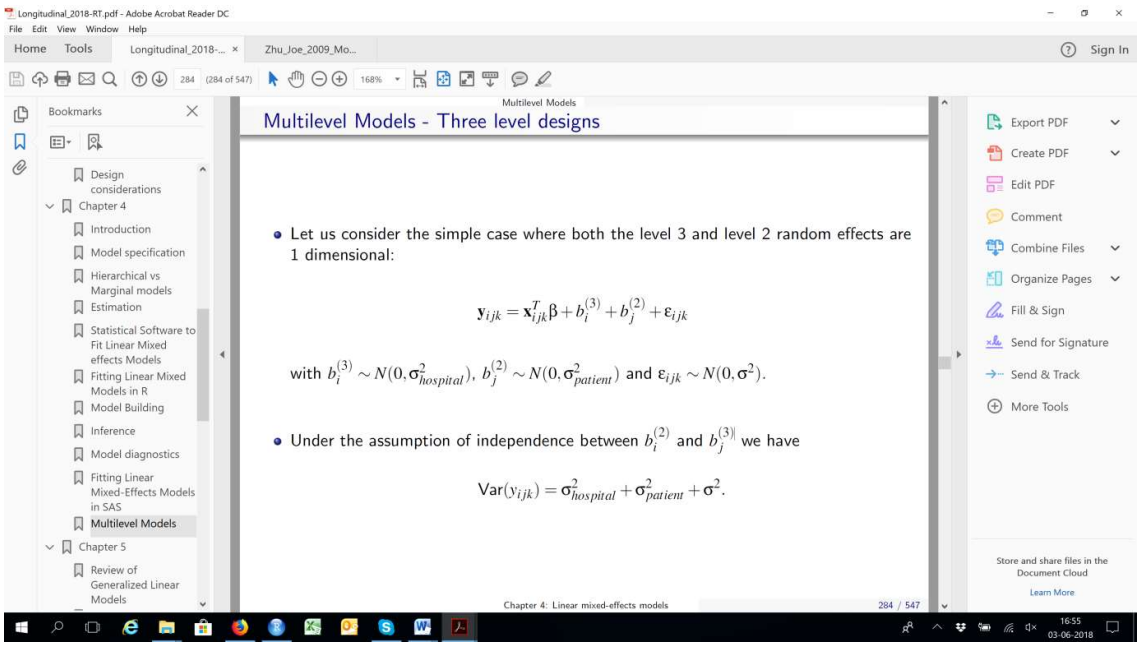
$$Z_i = \begin{bmatrix} 1 & 0 \\ 1 & 100 \\ 1 & 200 \end{bmatrix}$$

In **model 3** the original outcomes are used. Look at the R code above the output for model 3 at the end of this exercise and notice the specification of the random effects structure.

k. **[2.0]** How many random effects are assumed in this case? What is the distributional assumption made?

Solution:

It resembles the design of the lectures where we had patients nested within hospitals. Here we have visits nested with patients.



The outcomes are clustered by diaries/visits. It is likely that there is within diary/visit correlation, which is modelled by the second random effect. Here all days within a diary are assumed to be equally correlated, which is maybe too simple.

l. **[6.0]** What is the variance within visits and what is the correlation between days within the same visit?

Solution:

Within diary variance = 4.823792^2+3.188376^2+2.01315^2=37.48748. The correlation within diaries is (4.823792^2+3.188376^2)/(37.48748) = 0.8918901.

m. **[2.0]** The same data have been analysed using the Generalized Estimating Equations (GEE) approach where the fixed effects part of model 1 has been used to model the mean daily intake in time and an AR1 correlation matrix has been assumed to model the within infant correlations. Would you trust the results of this analysis? Explain and motivate your answers.

Solution:

The GEE will give valid results provided that sandwich estimator is used and missing data mechanism is MCAR.

After fitting all these models, the investigators realize that not all planned measurements have been collected because some infants have dropped out. The reason for dropout is that the daily intake for some infants was lower than normal and weight loss was observed. Therefore, they switched to another formula.

n. **[2.0]** What is the implied missing data mechanism? Are the results obtained under each of the models 1-3 valid regarding the missing data mechanism?

Solution:

The implied mechanism is missing at random because the missingness depends on the decreasing daily intake and thus an alternative has been proposed to the children. All analyses are valid under MAR as the model 1-3 are estimated using likelihood based approaches.

o. **[2.0]** The investigators are primarily interested in differences between the two groups in mean daily intake at week 4 of age of the infant. Assuming that the missing data mechanism is the one that holds under question p above, is it correct to apply a two samples t-test that uses only the data at this age of the infants? Motivate your answer.

Solution:

The implied mechanism is missing at random thus the data of week 4 is not a random sample of the population of infants and thus bias will arise if we use only a part of the data. Efficiency will also be lost.

Output exercise 2 – models 1, 2, and 3

Model 1, aggregated data

R code

```
library(nlme)
model.aggr <- lme(meany ~ meantime+ group + meantime:group,
                          random = ~ 1|id,
                          data = data.aggr2)
```

R output

```
summary(model.aggr)

Linear mixed-effects model fit by REML
 Data: data.aggr2
       AIC      BIC    logLik
  661.2938 676.6799 -324.6469

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    5.953944 2.232729

Fixed effects: meany ~ meantime + group + meantime:group
                   Value Std.Error DF  t-value p-value
(Intercept)    8.241900 1.5492797 96 5.319827  0.0000
meantime       1.084706 0.1901448 96 5.704631  0.0000
group          1.422030 2.2432394 96 0.633918  0.5276
meantime:group 0.397322 0.2812229 96 1.412836  0.1609

 Correlation:
               (Intr) meantm group
meantime       -0.826
group          -0.691  0.570
meantime:group  0.558 -0.676 -0.823

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-0.79783564 -0.23471420  0.01293777  0.20501943  0.72007711

Number of Observations: 100
Number of Groups: 100



anova(model.aggr)
               numDF denDF  F-value p-value
(Intercept)        1    96 736.2126  <.0001
meantime           1    96  79.8885  <.0001
group              1    96   9.9934  0.0021
meantime:group     1    96   1.9961  0.1609
```

Model 2, aggregated data

R code

```
model.2 <- lme(meany ~ meantime+ group + meantime:group,
                  random = ~ meantime|id,
                  data = data.aggr2)
```

R output

```
summary(model.2)

Linear mixed-effects model fit by REML
 Data: data.aggr2
      AIC      BIC    logLik
  665.291 685.8058 -324.6455

Random effects:
 Formula: ~meantime | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev      Corr
(Intercept) 6.394928962 (Intr)
meantime    0.005495021 -0.999
Residual    0.001195247

Fixed effects: meany ~ meantime + group + meantime:group
                   Value Std.Error DF  t-value p-value
(Intercept)    8.243785 1.5551147 96 5.301078  0.0000
meantime       1.084426 0.1901779 96 5.702167  0.0000
group          1.418286 2.2513706 96 0.629966  0.5302
meantime:group 0.397891 0.2812067 96 1.414940  0.1603
 Correlation:
               (Intr) meantm group
meantime       -0.827
group          -0.691  0.572
meantime:group  0.560 -0.676 -0.824

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-4.252642e-04 -1.251149e-04  6.848532e-06  1.084925e-04  3.812241e-04

Number of Observations: 100
Number of Groups: 100
```

Model 3, original data

R code

```
model.nested <- lme(y ~ time+ group + time:group,
                    random = ~ 1|id/visit,
                    data = data.)
```

R output

```
Linear mixed-effects model fit by REML
 Data: data.
      AIC      BIC    logLik
  7192.33 7229.466 -3589.165

Random effects:
 Formula: ~1 | id
        (Intercept)
StdDev:    4.823792

 Formula: ~1 | visit %in% id
        (Intercept) Residual
StdDev:    3.188376  2.01315

Fixed effects: y ~ time + group + time:group
               Value Std.Error   DF   t-value p-value
(Intercept) 8.926883 1.0437325 1240  8.552847  0.0000
time        0.995592 0.0761666 1240 13.071243  0.0000
group       0.437706 1.5101088   98  0.289851  0.7725
time:group  0.505000 0.1101777 1240  4.583505  0.0000
 Correlation:
            (Intr) time   group
time       -0.711
group      -0.691  0.491
time:group  0.491 -0.691 -0.706

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.11377282 -0.63113806 -0.01283971  0.63481790  3.06831980

Number of Observations: 1492
Number of Groups:
           id visit %in% id
          100           250
```