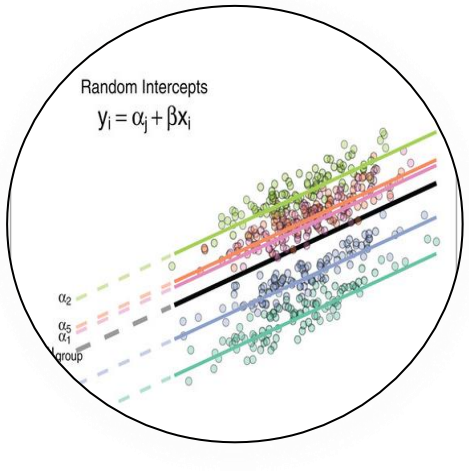
PhD course Mixed Linear Models

Session 5: Repeated measurements

Gerrit Gort (slides Bas Engel)
Biometris, Wageningen
University

June 10-11-12, 2025



WAGENINGEN UR
For quality of life

Repeated measurements

- *Multiple* measurements of the response variable on the *same* experimental unit
- Also known as *longitudinal data*
- Usually multiple measurements are made over a period of *time*, typical example: growth data
- Many books on this topic, e.g. *Analysis of Longitudinal Data* by Diggle, Liang & Zeger
- Closely related to *time-series analysis*
- Multiple measurements can also be made over *space*

- Measuring the same experimental unit repeatedly, causes *correlation* of the measurements from this experimental unit. By now, you can imagine that mixed models may be perfect framework to accommodate the correlations!

Repeated measurements in the past

- Historically, software like R and SAS has specific capabilities to model time-series data.
- Linear model routine like R's lm() or manova() have capabilities to handle multivariate responses, thereby including correlations among individual responses.
- Different procedures have been developed in past for analyzing repeated measurements, like:
  1) derived variable, like AUC, or slope of regression line
  2) separate analysis at each time point
  3) univariate analysis of variance ("split-plot", "Huyn-Feldt" conditions)
  4) univariate and multivariate analyses of time contrast variables
  5) mixed model methodology

  Nice overview in: Littell RC, Henry, PR, Ammerman, CB (1998) *Statistical Analysis of Repeated Measures Data Using SAS Procedures* Journal of Animal Science 76: 1216-1231

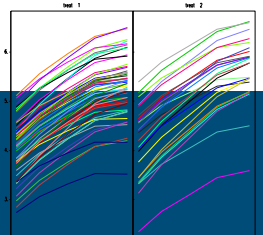
  (cited in WoS 879 times June 2010; 1025 times June 2011; 1132 times June 2012; 1252 times June 2013; 1373 times June 2014; 1465 times June 2015; 1598 times June 2016; 1666 times June 2017, 1758 times June 2018, 1873 times June 2019, 2003 times June 2020)
- We focus on 5) mixed model methodology, as you may expect. The split-plot type of analysis in 3) is also example of mixed model.

Repeated measures

- Classical design is completely randomized design with data collected on *equally spaced points*
- Experimental units are often called *subjects,* a heritage from human psychology
- Basic set-up: two factors, *treatments* and *time*
- Repeated measurement experiments are usually factorial experiments
- *Treatment = between subjects* factor, all measurements on same subject represent same treatment
- *Time = within subjects* factor

Example of repeated measurements data set

- Spruce trees
- Two treatments
  - ozone treated
  - control
- Four repeated measurements on size per tree
- We will see this dataset in exercise 11



Repeated measurements

- Research questions
  - How do treatment means *change over time*?
  - How do treatment differences change over time? (*treatment x time interaction*)
- The distinguishing characteristic of repeated measures studies is the *covariance structure* of the observations. Points closer together in time (space) are usually higher correlated than points further apart.
- Objective of repeated measurement study is to compare treatment means or treatment regression curves over time
- First, however, the covariance structure (especially within units or subjects) of the data needs to be assessed
- Modelling of the covariance structure is not a goal in itself, but is necessary to arrive at valid inferences for the fixed effects

Repeated measurements and split plots

- Treatment factor in a repeated measurements experiment corresponds to main plot factor in a split plot experiment
- Time factor corresponds to a subplot factor
- Difference: in a true split plot the levels of the subplot factor are randomly assigned and thus equal correlations among responses from subplots within a main plot may be a reasonable assumption
- So, when indeed the correlations between time points are more or less equal, a split plot analysis may be used for a repeated measurements analysis.

Mixed model $\underline{y} = X\beta + Z\underline{u} + \underline{e}$

- In repeated measurements analyses we concentrate especially on modelling the relations between the time points within subjects/units, so on $\underline{\boldsymbol{e}}$ and its variance-covariance matrix, R
- The R (residual) matrix in repeated measures models becomes *block diagonal*
- Some structures, like random coefficients, are specified through $Z\underline{u}$ part

Within unit (subject, whole plot) covariance structures

- Compound symmetry/ uniform correlation (default for Split Plot and Block designs)
- 4 timepoints

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_S^2 + \sigma_e^2 & & & \\ \sigma_S^2 & \sigma_S^2 + \sigma_e^2 & & \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_e^2 & \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_e^2 \end{bmatrix}$$

- σ^2_S is between subjects (whole plots) variance, σ^2_e is variance for individual time points (sub plots)

Alternative covariance structures in repeated measures studies

Autoregressive order one (AR1)

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

ρ is a correlation parameter typically between 0 and 1.

In AR1 model correlation decreases with increasing time difference

parameters: 2

Unstructured

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

In unstructured model, each time point has its own variance and each pair of time points has its own covariance.

parameters: 0.5*k*(k+1) = 10 here

Alternative covariance structures in repeated measures studies

Toeplitz

$$\Sigma = \begin{bmatrix} \sigma^2 & & & \\ \sigma_1 & \sigma^2 & & \\ \sigma_2 & \sigma_1 & \sigma^2 & \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

Observations at equal time lag have equal correlation

parameters: k = 4 here

Banded Toeplitz(q)

$$\Sigma = \begin{bmatrix} \sigma^2 & & & \\ \sigma_1 & \sigma^2 & & \\ 0 & \sigma_1 & \sigma^2 & \\ 0 & 0 & \sigma_1 & \sigma^2 \end{bmatrix}$$

Observations at equal time lag have equal correlation, but for time lag>(q-1) correlation=0

parameters: q = 2 here.

Within unit covariance structures, e.g.

- Compound symmetry / uniform correlation

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_S^2 + \sigma_e^2 & & & \\ \sigma_S^2 & \sigma_S^2 + \sigma_e^2 & & \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_e^2 & \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_e^2 \end{bmatrix}$$

```
gls(y ~ treat + time + treat:time,
 correlation = corCompSymm(form = ~ 1 | person))
```

Here we bypass the $Z\underline{u}$ part of mixed model

- Autoregressive order one

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

```
gls(y ~ treat + time + treat:time,
 correlation = corAR1(form = ~ 1 | person))
```

- Toeplitz; not available in the nlme package (as far as I know...)

$$\Sigma = \begin{bmatrix} \sigma^2 & & & \\ \sigma_1 & \sigma^2 & & \\ \sigma_2 & \sigma_1 & \sigma^2 & \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

Within unit covariance structures

- Toeplitz with two bands; not available in nlme (as far as I know...)

$$\Sigma = \begin{bmatrix} \sigma^2 & & & \\ \sigma_1 & \sigma^2 & & \\ 0 & \sigma_1 & \sigma^2 & \\ 0 & 0 & \sigma_1 & \sigma^2 \end{bmatrix}$$

- Unstructured, common variance

$$\Sigma = \begin{bmatrix} \sigma^2 & & & \\ \sigma_{21} & \sigma^2 & & \\ \sigma_{31} & \sigma_{32} & \sigma^2 & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma^2 \end{bmatrix}$$

```
gls(y ~ treat + time + treat:time,
 correlation = corSymm(form = ~ 1 | person) )
```

- Unstructured, different variances

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

```
gls(y ~ treat + time + treat:time,
 correlation = corSymm(form = ~ 1 | person),
 weights = varIdent(form= ~1 | time) )
```

Random coefficients model

$$\underline{y}_{ij} = \mu + \underline{a}_i + \underline{b}_i x_{ij} + \underline{e}_{ij}$$

$$\begin{pmatrix} \underline{a}_i \\ \underline{b}_i \end{pmatrix} = N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix} \right)$$

$$\underline{e}_{ij} = N(0, \sigma_e^2)$$

or

$$\underline{y}_{ij} = (\mu + \beta x_{ij}) + (\underline{a}_i^* + \underline{b}_i^* x_{ij}) + \underline{e}_{ij}$$

$$\begin{pmatrix} \underline{a}_i^* \\ \underline{b}_i^* \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix} \right)$$

$$\underline{e}_{ij} = N(0, \sigma_e^2)$$

```
lme(size ~ treat + time + treat:time, random= ~ 1 + time | person)

lmer(size ~ treat + time + treat:time + (1 + time | person))
```

- Peculiarities of random coefficients model:
  - variance is quadratic function of time!

$$var(\underline{y}_{ij}) = var\left((\underline{a}_i^* + \underline{b}_i^* x_{ij}) + \underline{e}_{ij}\right) = \sigma_a^2 + x_{ij}^2 \sigma_b^2 + 2 x_{ij} \sigma_{ab} + \sigma_e^2$$

  - covariance σ_{ab} between intercept and slope is needed, otherwise covariance structure will depend on time-scale (e.g. use x-60)

Modelling covariance structure

- Usually not goal in itself, but necessary for valid inferences of fixed effects

- Compare models:
  - *-2 Residual Log Likelihood* : smaller is better

    smaller (nested) model **always** smaller log likelihood, hence larger -2 log likelihood

    LR hypothesis test by comparison of nested models
  - ***Not*** usable for non-nested models!
  - Keep fixed part of model constant, if REML is used!

- Penalized log likelihoods:
  - $AIC = -2\ Res\ LL + 2\ k$ (k = number of parameters); the smaller the better
  - $AICC = -2\ Res\ LL + 2\ k\ \ n/(n-k-1)$ Corrected AIC, with small-sample-size correction
  - $BIC = -2\ Res\ LL + k\ \ log(n)$ Bayesian Information Criterion = Schwarz' criterion

Exercise 11: Random coefficients model

In this exercise we look at some data of one of the participants (Martin de Graaf) in a PhD course some years ago. In the study the handling time, that it takes a predator fish (species *Acutirostris*) to swallow a prey fish, is measured. Prey fishes of different species and of different sizes are used. Each predator fish is given prey fishes of different sizes (but of a single species: Garra, Humilis, Niloticu or Tanapela). Numbers of prey fishes per predator varies from 2 to 19.

a) Read data into the system. In total 23 predators, which swallow multiple prey fishes.

b) Make plot of all 23 individual regression lines, regressing swallowing time on size of prey fish.

c) Fit a random coefficients model, but not with time as regressor, but prey fish size. The fixed part part of the model allows 4 separate regression lines for predator fishes swallowing 4 types of prey fish. The individual predators have regression lines deviating from the 4 population lines with random intercept deviations and random slope deviations.

Exercise 11: Random coefficients model, continued

d) May be the random coefficient model is too complex. Try a model with random intercepts only. Compare the AIC's, but also use a likelihood ratio test (as we have nested models here). What is the difference in number of parameters between the two models? You may use PQRS to calculate the P-value.

[Besides: there is a problem with this hypothesis testing approach, because under the null hypothesis the variance of the slope takes a value on the boundary (0). Therefore, the distribution of the test statistic under H_0 is some mixture of chi-square distributions.]

d) May be the random intercept model is too complex as well. Therefore, fit an ordinary linear model and compare AIC's and use a likelihood ratio test.

e) The bottom part of the program file contains some code to make a plot of the results of the random coefficient model (though the model is overly complex). Notice how strong the regression lines are shrunken towards to the population regression lines, compared to the plot made in part b).

Exercise 12: Repeated measures of spruce trees

- In this exercise you find the famous Sitka spruce data, table 1.2 from Diggle, Liang and Zeger, Analysis of longitudinal data, Oxford Science Publications. Size was measured on 79 trees in 1988 at 152, 174, 201, 227 and 258 days after January 1. Trees 1 to 54 were exposed to ozone enriched atmosphere, trees 55 till 79 formed the controls (normal atmosphere). The question we want to answer is whether increased ozone concentrations (global heating) influence the growth of trees.

- File exercise12.r contains an analysis. Four models are fitted to the data:
  - repeated measurements with compound symmetry for variance covariance matrix within trees
  - as 1, but using an autoregressive process with lag 1
  - as 1, but using an unstructured variance covariance matrix
  - a random coefficients model.

  Compare the models and choose a best model. For the choice of best model, use Akaike's Information Criterion (AIC= -2 log likelihood + 2d with d number of parameters), smaller is better.

Exercise 12: Repeated measures of spruce trees, continued

- Next compare the 4 models with respect to
  - The F values for the fixed terms of ozone, time and their interaction.
  - The estimates for the fixed effects and their standard errors. What is the equation of the fitted lines for the ozone and control treatment?
  - The fitted variance covariance matrices within the trees.
  - The fitted correlation matrices within the trees..

Graphical checks of model fit

- Graphical checks in Mixed Models not as frequently done as in Linear Models
  - mainly due to lack of methodology, but things are coming...
  - but certainly adequacy of model needs to be checked in Mixed Models as well: e.g. presence of outliers, heterogeneous variance, need for data transformation
- Different types of residuals:
  - marginal residuals $r_m = \underline{y} - X\hat{\beta}$
  - conditional residuals $r_c = \underline{y} - X\hat{\beta} - Z\underline{\hat{u}}$
  - both can be (internally) *studentized* , i.e. divide by estimated standard error
  - both can be *externally studentized*, i.e. without using observation *i* itself
  - both can be of Pearson type, i.e. dividing by std(y_i)
  - scaled marginal residuals (by inverse Cholesky-root of V)
- Marginal and conditional residuals have different properties
  - marginal residuals have expectation 0, but do not necessarily sum to 0
- Try out plot() function for lme-object.
- Beyond scope of present course

Heterogeneous variances and correlation in space

- Heterogeneous variants of some covariance types may be interesting.
- In SAS these would be called:
  - ARH(1) = heterogeneous AR(1) → (i,j)th element is $\sigma_i \sigma_j \rho^{|i-j|}$
  - CSH = heterogeneous CS → (i,j)th element is $\sigma_i \sigma_j \left[\rho 1(i \neq j) + 1(i = j)\right]$
  - TOEPH = heterogeneous TOEP → (i,j)th element is $\sigma_i \sigma_j \rho_{|i-j|}$
- In R package `nlme` obtained through `weights` and `varIdent`.

- Mixed models can also be used to model spatial data. Some spatial covariance structures are available within package `nlme`.

- New R packages are appearing all the time like `spaMM` for inference in mixed models, in particular for spatial (GL)MM's, `merlin`, `glmmTMB`, ...