

## Index of Lecture 12a2: Linear mixed models

Page	Title
1	Practical information
2	Mixed models for continuous data, incl. variance component estimation
3	Conceptual example (review)
4	Mixed model for hierarchical data
5	Hierarchical data structure
6	Random effects
7	Somatic cell count datasets
8	Mixed model for somatic cell count data
9	Reunion Island study
10	Variance components
11	Estimation in linear mixed models
12	Statistical inference in linear mixed models
13	Statistical inference (continued)

## PRACTICAL INFORMATION

### Today's lecture:

- linear mixed models — follow-up to Javier's introduction to clustered data:
  - \* introduction to analysis focusing on main principles,
  - \* more details in: regular version of VHM 802 and multilevel summer course (not offered this year),
  - \* main focus: analysis using likelihood-based estimation (all software packages<sup>1</sup>),
- **textbook coverage** (for lecture): VER2/MER, Sections 21.1-2 and parts of 21.5.

### Schedule:

- **review** session for lecture (with discussion, and potential software demonstrations): Tuesday next week (April 6),
- **practice problem** for lecture and VER Chapter 20 (and 21) using dataset ap2\_intro: Questions 1-4 (Question 5 is optional), also to be reviewed Tuesday next week,
- **next lecture** (recorded video): Thursday this week, on discrete data and alternative methods.

---

<sup>1</sup> Minitab 18+: Stat-ANOVA-Mixed Effects menu; R: nlme and lme4 libraries.

## MIXED MODELS FOR CONTINUOUS DATA, INCL. VARIANCE COMPONENT ESTIMATION

### Synthesis:

- mixed models **extend ordinary linear models** (regression and ANOVA) to take into account “clustering”.

### Contents:

- **introduction** to mixed models and modelling
  - \* theory (gently), notation and practice,
  - \* brief overview of main modelling steps,
- Stata computer demonstrations.

### Terminology and relationships:

- mixed  
○ random effects  
○ variance component } models – the same,
- “mixed”  $\sim$  containing both fixed and random effects,
- multi-level  
○ hierarchical } models – the same, and **special type of mixed models**,
- variance components are mathematical constructs used in mixed models.

## CONCEPTUAL EXAMPLE (REVIEW)

Consider the following problem:

- study of risk factors for (high) somatic cell counts (e.g., as a crude indicator of mastitis),
- **one recording** (for simplicity) of the cell count in a milk sample **from each cow**; in total,  $n$  cows,
- additional recordings of **explanatory variables for each cow**, such as lactation stage (days in milk), age, breed,...
- also **explanatory variables at the herd level**, e.g. housing type, herd size,...
- **linear model**:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad \text{or} \quad Y = X'\beta + \varepsilon$$

where

- \*  $Y_i$  = (natural) log somatic cell count for cow  $i$ ,  $i = 1, \dots, n$ ,
- \*  $x_{ri}$ 's contain values of the explanatory variables,<sup>2</sup>
- \*  $(\beta_0), \beta_1, \dots, \beta_k$  are regression coefficients for  $x$ 's,
- \*  $\varepsilon_i$  = error term  $\sim N(0, \sigma^2)$ .

---

<sup>2</sup> In this notation (from the VER2 textbook), we use  $x_{ri}$  instead of the usual  $x_{ir}$ ;  $X' = (x_{ri})_{ir}$  is the  $n \times (k+1)$  design matrix, including as  $(x_{0i})$  a column of 1's.

## MIXED MODEL FOR HIERARCHICAL DATA

**Simplest case**  $\sim$  extended cell count example, for measurements on cows in several herds,

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + u_j + \varepsilon_{ij}, \quad \text{or} \quad Y = X'\beta + u + \varepsilon$$

where

- $Y_{ij}$  = log somatic cell count for cow  $i$  in herd  $j$ ,
- $u_j$  = random  $j^{\text{th}}$  herd effect  $\sim N(0, \sigma_h^2)$ ,
- $\sigma_h$  = scale of random herd effects, interpretable as the amount of random variation in log-scc between herds;
  - \* e.g., 95% of herds expected within  $0 \pm 1.96 \sigma_h$ ,
- $i$  = cow number,  $j$  = herd number.

**Definitions** (non-Bayesian terminology):

- “**random**” effect: a model term (right hand side) which is a random variable (often not counting the error term  $\varepsilon$ ),
- “**fixed**” effect: modelled by non-stochastic parameters ( $\beta$ 's, often not counting the intercept  $\beta_0$ ),
- **mixed model**: both fixed and random effects.

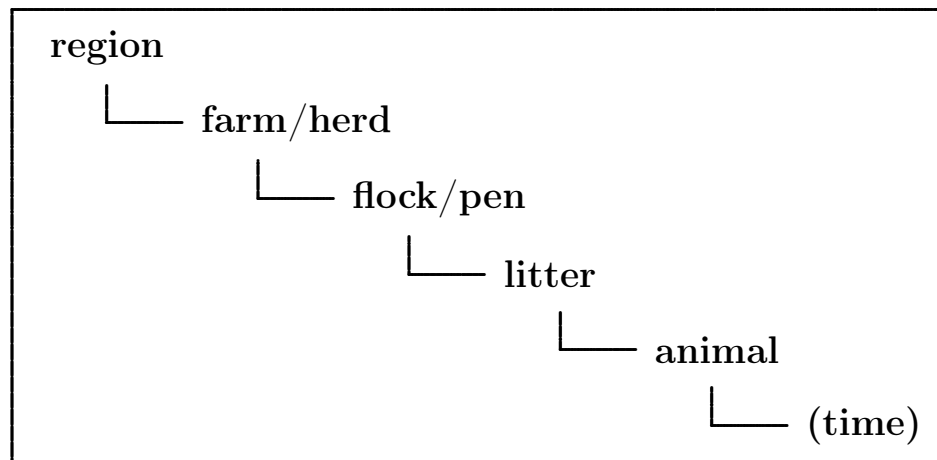
## HIERARCHICAL DATA STRUCTURE

A **hierarchical data structure**, typically implying both of the following:

- observations grouped at different levels,
- factors (e.g. treatments) reside (or are applied) at different levels.

may induce **clustering** in the data, that is, some observations are more alike than others, or put in another way: the **observations are no longer independent**.

**Typical example** from  
veterinary epidemiology:



**Note:** “time” as a bottom level  $\sim$  longitudinal data / repeated measures on the same animal, and raises some additional modelling issues (next lecture).

**Random effects models for hierarchical data:**

- insert random effect(s) for each hierarchical level (above the bottom level).

## RANDOM EFFECTS

- the only new concept in mixed models,
- enable a **separation (and quantification)** of variation at different levels in the data,
- enable **correct analysis of predictors at different levels** within the same model,
- involve **additional assumption(s)** of normal distribution and variance homogeneity.

**Motivations for random effects** (decreasing importance):

- hierarchical data structure:
  - \* **rule**: insert a random effect for each hierarchical level above the bottom level (exceptions: 12a2L–13),
- correct analysis of treatments allocated to **larger exper. units** (“split-plot” idea<sup>3</sup>),
- factor where **interest is in the variation between units** (within a level) rather than specific units in study:
  - \* units may be randomly selected,
  - \* units should **represent “population”** – to which the conclusions from the study may be generalized,
- avoid many (nuisance<sup>4</sup>) parameters in model/estimation.

---

<sup>3</sup> Split-plot designs are experimental designs where treatment factors are applied to units of different sizes; discussed in detail in regular version of VHM 802, and in the GO textbook.

<sup>4</sup> Of no or little intrinsic interest.

## SOMATIC CELL COUNT DATASETS

scc\_40 — a real somatic cell count dataset:

A subset comprising 40 herds from a large dataset collected in 1993-94 by Jens Agger and co-workers including about 2150 Danish herds and 150 000 cows followed throughout one lactation. The data contain approximately monthly milk records plus information collected through herd questionnaires.

Variable	Description	Values
herdid	herd id	1 – 40
cowid	cow id	1 – 2178
test	approximate month of lactation	0 – 10
t_lnscc	natural log scc (in 1000s) on test day	2.3 – 9.2
t_dim	days in milk on test day	10 – 305
t_season	season of test day	1 – 4 (1 = Jan-Mar, etc.)
c_heifer	parity of cow	0/1 (1 = heifer)
h_size	average herd size	10.3 – 101.5
t_ecm	energy-corrected <sup>1</sup> milk yield	2.2 – 68.5

<sup>1</sup> computed by the formula:

$$\text{ecm} = \text{kgmilk}(0.383 \text{ fatpct} + 0.242 \text{ proteinpct} + 0.7832)/3.14.$$

**Subdataset** scc40\_2level: only first observation per lactation included  
 ⇒ one observation per cow.

## MIXED MODEL FOR SOMATIC CELL COUNT DATA

- **data:** 2-level somatic cell count data (scc40\_2level),
- **outcome:** log somatic cell count (t\_lnscc),
- **fixed effects:** season (categorical), dim, heifer, hsize (all quantitative),
- **random effects:** herds (note: only 1 observation per cow).

```
. mixed t_lnscc h_size c_heifer i.t_season t_dim || herdid:, reml
```

```
Mixed-effects REML regression          Number of obs      =      2178
Group variable: herdid                 Number of groups   =        40
                                       Obs per group: min =        12
                                       avg =          54.5
                                       max =          105
                                       Wald chi2(6)       =      244.36
Log restricted-likelihood = -3624.9622  Prob > chi2        =      0.0000
```

```
-----+-----
      t_lnscc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      h_size |   .0040837   .0037726     1.08   0.279   - .0033105   .0114778
...
      _cons |   4.641202   .1974215    23.51   0.000    4.254263    5.028141
-----+-----
```

```
Random-effects Parameters |   Estimate   Std. Err.     [95% Conf. Interval]
-----+-----
herdid: Identity         |
      var(_cons) |   .1491533   .0436191     .0840821   .2645832
-----+-----
      var(Residual) |   1.557228   .0477206     1.466451   1.653625
-----+-----
```

```
LR test vs. linear regression: chibar2(01) =    97.01 Prob >= chibar2 = 0.0000
```

## REUNION ISLAND STUDY

- carried out 1993-1996 on Reunion Island (Emmanuel Tillard, CIRAD),
- data analyzed 1996-2000 (with a helping hand from Ian Dohoo), and results communicated to the cattle industry and published 1999-2001.

**Objective:** identify the factors and levels at which most of the variability in reproductive performance resides ... where interventions are likely to have the most effect.

**Reproductive performance in cows** measured by time from calving to conception, which is composed of

- time from calving to first service,
- conception from first service (yes/no),
- if no, time from first service to conception.

**Data size and structure:**

Level	Number of units	Per unit at level above	
		Average	Range
Region	5	—	—
Herd	50	10.0	3–16
Cow	1575	31.5	8–105
Lactation	3027	1.9	1–5

- no cow movements between herds (~ strict hierarchical structure).

## VARIANCE COMPONENTS

Always a **decomposition of the total variation**:

total variation = fixed effects variation + random variation,<sup>5</sup>

Additionally in random effects models — a **decomposition of the random variation**.

**2-level model** — cell count example (herds—cows):

- $\text{Var}(Y_{ij}) = \sigma^2 + \sigma_h^2$  (= total unexplained random variation),
- variance components  $\sigma^2$  and  $\sigma_h^2$ ,
- of the total random variation,  $\sigma_h^2/(\sigma^2 + \sigma_h^2)$  resides at the herd level, and the rest at the cow level; the former is also termed a **variance partition coefficient** (VPC),
- $\text{ICC}^6 = \sigma_h^2/(\sigma^2 + \sigma_h^2)$ , often denoted  $\rho$  (as a correlation coefficient).

**Multilevel models** — Reunion example (herds—cows—lactations):

- $\text{Var}(Y_{ij}) = \sigma^2 + \sigma_c^2 + \sigma_h^2$ ,
- **proportions of variance** (VPCs) at different levels in the obvious way,
- **two ICCs**:<sup>7</sup> 
$$\left\{ \begin{array}{l} \text{lactations of same cow : } (\sigma_c^2 + \sigma_h^2)/(\sigma^2 + \sigma_c^2 + \sigma_h^2), \\ \text{lactations in same herd : } \sigma_h^2/(\sigma^2 + \sigma_c^2 + \sigma_h^2) \end{array} \right. ,$$

<sup>5</sup> Least squares decomposition:  $\sum(Y_i - \bar{Y})^2 = \sum((X'\hat{\beta})_i - \bar{Y})^2 + \sum(Y_i - (X'\hat{\beta})_i)^2$

<sup>6</sup> Intra-class (or -cluster) correlation coefficient: the correlation between two observations in the same class/cluster. Alternative **interpretations of clustering** as: variation between clusters, or correlation within clusters.

<sup>7</sup> **General formula**: sum of variance components of **common random effects** divided by sum of all variance components.

## ESTIMATION IN LINEAR MIXED MODELS

“Likelihood”-based estimation assuming normal distributions for all random terms:

- REML (restricted maximum likelihood) or (full) ML:
  - \* theoretical properties differ slightly (REML unbiased, ML less variance),
  - \* in practice only minor differences, unless the number of units at a level is small,
  - \* REML estimates agree with ANOVA-type<sup>8</sup> estimates for balanced data,  
— choice between REML and ML is “a matter of taste” (but keep consistent),
- iterative, numerically robust algorithms,
- performs well for both balanced and unbalanced data,
- available in many statistical software packages, however however
  - \* different modelling flexibility,
  - \* different ability to handle large data structures,
- should give the same estimates from different software packages, up to estimation accuracy, despite minor differences in implementations.

---

<sup>8</sup> A classical statistical method for variance component models relies on the ANOVA-table, and constructs estimates and test statistics from the MS-column; performs well only for almost balanced data; discussed in (the standard) VHM 802 course and the GO textbook.

## STATISTICAL INFERENCE IN LINEAR MIXED MODELS

Tests and confidence intervals — approximate and assuming normal distributions for all random terms:

- **Wald statistics** for fixed effects: based on standard errors and estimated correlations between estimates,
  - \* **95% confidence intervals**:  $\hat{\beta}_r \pm 1.96 \times \text{SE}(\hat{\beta}_r)$ ,  
(better to use  $t^* \sim t$ -distribution with suitable df<sup>9</sup>, but similar if df is large),
  - \* simple to compute, for fixed effect parameters usually ok,
- **likelihood-ratio tests**, based on optimal values of likelihood function (more precisely, differences of  $-2 \log L$ ):<sup>10</sup>

$$G^2 = 2(\log L_{\text{full}} - \log L_{\text{red}}) \sim \chi^2(\text{df}),$$

where df = number of parameters being tested equal to zero,

- \* the **only appropriate test** for **variance parameters**,<sup>11</sup>
- **confidence intervals for variance parameters**: not easy, and usually ok to present Stata's approximate intervals from Wald-type procedure (but inference from these can be misleading).

---

<sup>9</sup> In Stata 14+, options `dfmethod(satterthwaite)` or `dfmethod(kroger)`; same options in Minitab 18+.

<sup>10</sup> Caution for fixed effects parameters and REML estimation: beware to **not use** the restricted likelihood. Stata's `lrtest` command gives a warning note.

<sup>11</sup> Note: the *P*-value should be **half** the value from  $\chi^2(\text{df})$  when testing  $H_0 : \sigma_h^2 = 0$ . (Stata does this per default.)

## STATISTICAL INFERENCE (CONTINUED)

### Model-building guidelines:

- **fixed effects**: similar to linear models,<sup>12</sup>
- **random effects**: generally one for each hierarchical level, but some exceptions:
  - \* fixed effects possible, if **number of units is small** and/or **unrepresentative** of a population, and no predictor has variation at that level — this typically occurs at the highest level,
  - \* level may need to be omitted if only very few replications present in the data.

### Model checking:

- now model assumptions (and potential violations) at multiple levels,
- **predictions, residuals and diagnostics at multiple levels**,
- softwares differ in which of these statistics are available; Stata gives easy access to
  - \* lowest level residuals (however, not standardized well) + fitted values  
⇒ usual model checking at lowest level,
  - \* predicted random effects (“BLUPs”) at higher levels ⇒ normality checks can be done easily, and plots against predicted values with some coding,
- **model checking is an informal, exploratory process.**

<sup>12</sup> **Recommendation**: fixed effects model-building should use models with all relevant random effects.