# Index of Lecture 10b

# Practical Information

Today's lecture:

- introduction to clustered data (Javier),
- linear mixed models (Henrik/these notes):
  - introduction focusing on main principles,
  - more details in
    - VHM 802 (not offered this year),
    - multilevel summer course (June 2018; more information at `http://cver.upei.ca`).

Textbook (VER2/MER) reading:

- Clustered data: Sections 20.1-3 with some skippings,
- Linear mixed models: Sections 21.1-2 and parts of 21.5,

Home work for Tuesday next week (review session):

- Sample problem for VER Chapter 20 (and 21) using dataset `ap2_intro`: Questions 1-4 (Question 5 is optional).

# 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" ∼ 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

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,
- *linear model*:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \qquad (1) \\ &\text{or} \\ Y &= X'\beta + \varepsilon\,, \end{aligned}$$

  where

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

[1] 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 ~ extended cell count example:
assume measurements on cows in several herds,

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_k x_{kij} + u_j + \varepsilon_{ij}, \qquad (2) \\ &\text{or} \\ Y &= X'\beta + u + \varepsilon\,, \end{aligned}$$

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.

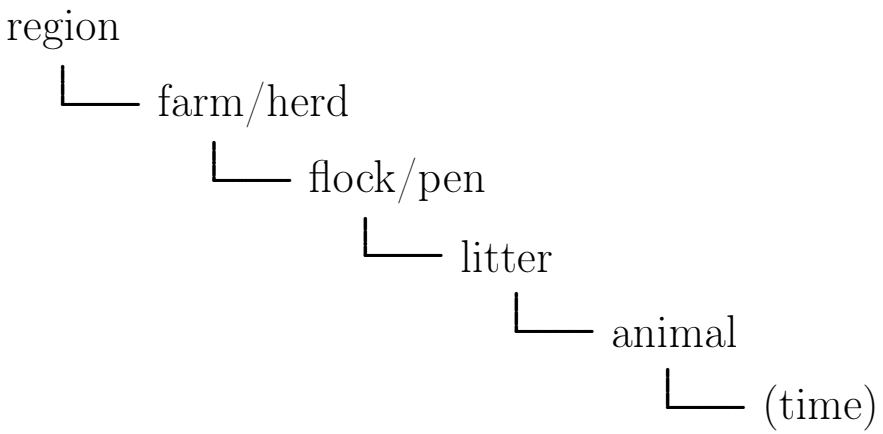
# Hierarchical Data Structure

A hierarchical data structure:

- observations grouped at different levels,
- treatments 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.

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: L10b–13),
- correct analysis of treatments allocated to larger experimental units ("split-plot" idea[2]),
- 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[3]) parameters in model/estimation.

[2] Split-plot designs are experimental designs where treatment factors are applied to units of different sizes; discussed in detail in VHM 802.

[3] Of no or little intrinsic interest.

# Somatic Cell Count Datasets

`scc_40` – a real somatic cell count dataset:
A subset including 40 herds from a large dataset collected in 1993-94 by J. Agger and co-workers including about 2150 Danish herds and 150 000 cows followed throughout one lactation. The data contain approx. monthly milk records plus information collected through herd questionnaires.

| Variable | Description | Values |
|---|---|---|
| `herdid` | herd id | $1-40$ |
| `cowid` | cow id | $1-2178$ |
| `test` | approx. 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[1] milk | |
| | yield | $2.2-68.5$ |

[1] computed by the formula:

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

Subdataset `scc40_2level`:
only first observation per lactation included
$\Rightarrow$ 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 (categ.), dim, heifer, hsize (all cont.)
- random effects: herds (because only 1 obs. 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
    c_heifer |  -.7367168   .0554447   -13.29   0.000    -.8453863   -.6280472
             |
    t_season |
    apr-jun  |   .1609431   .0906574     1.78   0.076    -.0167422    .3386285
    jul-sep  |   .0015031   .0863774     0.02   0.986    -.1677935    .1707997
    oct-dec  |   .0014582   .0918936     0.02   0.987    -.1786499    .1815663
             |
       t_dim |   .0027731   .0004991     5.56   0.000     .0017949    .0037513
       _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,
- data analyzed 1996-2000, 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 | 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 ($\sim$ strict hierarchical structure).

# Variance Components

Always a decomposition of the total variation:

total variation = fixed effects var. + random var.[4]

Additionally in random effects models:
− a decomposition of the random variation.

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

- $\mathrm{Var}(Y_{ij}) = \sigma^2 + \sigma_h^2$ (= total unexplained random var.),
- 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,[5]
- ICC[6] $= \sigma_h^2/(\sigma^2+\sigma_h^2)$, often denoted $\rho$ (as a corr. coef.).

Multilevel models − Reunion ex. (herds−cows−lact.):

- $\mathrm{Var}(Y_{ij}) = \sigma^2 + \sigma_c^2 + \sigma_h^2$,
- proportions of variance at diff. levels in the obvious way,
- two ICCs:

  lactations of same cow : $(\sigma_c^2+\sigma_h^2)/(\sigma^2+\sigma_c^2+\sigma_h^2)$,
  lactations in same herd : $\sigma_h^2/(\sigma^2+\sigma_c^2+\sigma_h^2)$,

  general formula: sum of variance components of common random effects divided by sum of all variance comp.

[4] Least squares decomposition: $\sum(Y_i-\bar{Y})^2 = \sum((X'\hat{\beta})_i-\bar{Y})^2+\sum(Y_i-(X'\hat{\beta})_i)^2$

[5] Also termed variance partition coefficient (VPC).

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

# 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[7] estimates for balanced data,
- iterative, numerically robust algorithms,
- performs well for both balanced and unbalanced data,
- available in many statistical software packages (but not in Minitab), 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.

[7] 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.

## 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 $t$-distribution percentile with suitable df[8], 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 (differences of $-2 \log L$):[9]

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

  where df = no. of parameters being tested equal to zero,
  - the only appropriate test for variance parameters, [10]
- confidence intervals for variance parameters: not easy; usually ok to present approximate intervals in Stata from Wald-type procedure (but inference from these can be misleading).

[8] In Stata 14+, options `dtmethod(satterthwaite)` or `dfmethod(kroger)`.

[9] Caution for fixed parameters and REML estimation: beware to *not* use the restricted likelihood. Stata's `lrtest` command gives a warning note.

[10] Note: the $P$-value should be *half* the value from $\chi^2(\text{df})$ when testing $\text{H}_0 : \sigma_h^2 = 0$. (Stata does this per default.)

# Statistical Inference (cont.)

Model-building guidelines:

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

Model checking:

- model assumptions (and potential violations) at multiple levels,
- predictions, residuals and diagnostics at multiple levels,
- software differences in the accessibility of these statistics; Stata gives easy access to
  - lowest level residuals (not standardized well) + fitted values $\Rightarrow$ usual model checking at lowest level,
  - predicted random effects at higher levels $\Rightarrow$ normality checks can be done easily, and plots against predicted values with some work,
- model checking is an informal, exploratory process.

# Stata Do-file (Selection)

```
use scc40_2level, clear
mixed t_lnscc h_size c_heifer i.t_season t_dim || herdid:, reml
di 1.96*sqrt(.1491533)
di .1491533/(.1491533+1.557228)  /* ICC and VPC */
testparm i.t_season  /* Wald test for season */

* Reunion Island data, 3-level model for (natural) log cfs
use reu_cfs, clear
mixed lncfs i.heifer || herd: || cow:, reml
* ICC: same herd
di .0145193/(.0145193+.0200572+.1341146)
* ICC: same cow
di (.0145193+.0200572)/(.0145193+.0200572+.1341146)
* ICC calculation by Stata
estat icc
* model checking
predict res_lact, rstandard
summarize res_lact, d
qnorm res_lact
histogram res_lact
swilk res_lact
predict fitted, fit /* includes all random effects */
scatter res_lact fitted
* predicted random effects at higher levels
predict ref*, reffects
bysort cow: generate within_c=_n
bysort herd: generate within_h=_n
* herd-level random effects
summarize ref1 if within_h==1, d
qnorm ref1 if within_h==1
swilk ref1 if within_h==1
* cow-level random effects
summarize ref2 if within_c==1, d
qnorm ref2 if within_c==1
swilk ref2 if within_c==1
```