MLM 2021 summary

# Take Home Message

## Module 1 - Data Structures and Model Building

***What Can We Learn?***

* The important steps in the analysis through MLM include
  + the *understanding* of the data *structure* for a first step in model building,
  + the transcription of this understanding into a *statistical model* with associated *hypotheses*,
  + a *descriptive analysis* to investigate how these *hypotheses* hold or not,
  + a potential new *model building* that take into account what has been observed.
* The important elements in MLM are
  + the *fixed* and *random* effects,
  + the *within* and *between* subjects factors or regressors,
  + the possible *nesting* effects or *hierarchical* effects,
  + the possible interactions between fixed, random or both types of effects,
  + if all the considered effects can be estimated or if they are *confounded* with the *residual error*.

*Supplementary material* - (Descriptive Analysis for MLM with R):

Key ideas: boxplots, interaction plots, grouped data graphics, correlation structure graphic

## Module 2 - Formulation of Mixed Linear Models

***What Can We Learn?***

* A proper and unified MLM formulation is important for several reasons:
  + communication between scientists,
  + practical derivation of model's moments,
  + helpful for setting up simulation studies (see later).
* Learning how to formulate a MLM requires basic knowledge about *linear algebra*.
* There are different levels of MLM formulation, which are suitable for different purposes:
  + At the subject/cluster level: better to derive likelihood functions and its derivatives,
  + At the global level: for simulations and resampling methods (see later).

*Supplementary material* - (Linear Algebra: Useful Properties, Matrix Formulation of MLM):

Key ideas: view data, define model (single formula expression), consider single formula expression to develop (all) sampled individual formulas, convert to linear algebraic expression, -> leads to correlation matrix and use R code to produce graphic of correlation structure

## Module 3 - Estimation Methods for MLM

***What Can We Learn?***

* Estimation of MLM includes:
  + estimation of the fixed effects parameters ( β ),
  + estimation of the random effects variances (*variance components*).
* For estimating the variances, one can use
  + the MLE, but it is biased (in small samples),
  + the REML which is unbiased, but inference is not accurate (see Module 4).
* For estimating the fixed effects, one can use the MLE with
  + the variances estimated using the MLE, or
  + the variances estimated using the REML.
* The suitability of each estimator concerns inference (see Module 4).
* The fixed effects interpretation depends on the contrast matrix (see Module 4).

*Supplementary material* - (Fitting with R, Fitting with R:Examples)

Key ideas: contrast matrices, fixed effects, random effects, single-level, crossed, multi-level (nested)

Mixed linear models can be estimated by maximum likelihood (MLE) with underestimated variances in small samples or with restricted maximum likelihood (REML) with unbiased variances but with estimations that are not invariant to the choice of the contrast matrix for fixed effects. Thus, if we have a special interest on variances (to compare for example the residual variance and the variance due to a random effect), we shall use the REML, but if are interested on statistics that depend on the fixed parameters (inference for β, comparison of models), we shall use the estimations of maximum likelihood (MLE).

(R functions and packages: lme, lmer, nlme, lmerTest, lme4)

## Module 4 - Testing Hypotheses with MLM

***What Can We Learn?***

* Testing Hypotheses with MLM (and LM) becomes involving when there are factors in the model.
* Testing the fixed effects coefficients (β) implies a careful consideration of the choice of the contrast matrix.
* Testing for factors' effects depend on several features for the setting:
  + when the data are balanced and if the contrast matrix is chosen such as to provide an orthogonal design matrix X, then the different methods are equivalent.
  + in the other cases, there are notable differences, which correspond to different H0.
  + In these cases, there is no unanimity about which type of SS one should choose, but there seems to be a preference for the type III SS.
* Going further, when testing separately several hypotheses, one should be aware of the effects of taking decisions on the FWER.
* Testing procedures (corrections) are available for correcting the testing procedures for single hypothesis testing to control the FWER.
* These are particularly useful when testing MC, i.e. when comparing the differences in the mean effects of all the combinations of the levels of one factor.
* When there are interactions between factors, the number of possible comparisons can become very large, and the tests for MC are less powerful.

*Supplementary material* - (Testing With MLM in R)

In this document, we show how to use R to obtain the necessary information for testing hypotheses related to MLM. Testing single hypotheses is easily obtained, but the conclusions depend on the type of contrast that is used. For testing multiple hypotheses (i.e. for factors), the methods between packages vary, which can be very confusing.

Inference for single parameter: The interpretation of the tests depend on the chosen contrast for the factors. In the vast majority of the literature, a contrast is defined as an estimable (parametric) function such as a linear combination whose parameters sum to zero.

In software such as R and SAS, what is called a “contrast matrix” is not always similar to the notion of a (parametric) function contrast whose parameters always sum to zero. It is rather a full-rank matrix, whose rows are estimable functions, i.e. such that XTX is invertible. In both cases they are estimable functions of the model’s parameters.

In R, there are built in contrast matrices that can be used, but one can also build specific contrast matrices. The available contrast matrices in R are the treatment contrasts, the sum contrasts and the Helmert contrasts.

You can build any contrast of your choice as long as it is an estimable parametric functional.

Inference for multiple model parameters: With MLM (and LM), an interesting test occurs when the fixed effects part of the model contains factors (i.e. nominal variables). The aim is then to test if a factor, as a whole), has a significant effect on the response. This is different than testing single hypotheses (i.e. contrasts), since not all combinations of mean effects are tested, and it may occur that none of the fixed effects parameters is significant, while the corresponding factor has a significant effect on the response.

Traditionally, testing for factors goes through the ANalysis Of VAriance (ANOVA). ANOVA is based on the law of total variance, where the observed variance in the response is partitioned into components attributable to different sources of variation (i.e. factors and residual error). The variance decomposition is made through the Sum of Squares (SS) of the different sources of variability.

When the design is not balanced, e.g. when not all participants are measured in the same conditions (crossing of the factors’ levels), and when the design matrix (e.g. contrast matrix) is not orthogonal, then the marginal SS for each factor does not sum up to the total variance. In these cases, alternative decompositions are proposed.

There are three main approaches (or types) to do the ANOVA, which boils down to expressing the hypothesis testing procedures in three different ways: type I, II and III ANOVAs. The notation was introduced by SAS to define the type of sums of squares to be computed based on the expression of the full model (MF) and the restricted model (MR) in the hypothesis testing procedures.

The family-wise error rate relates to the inflated type-I error that occurs when conducting multiple hypothesis tests where each might potentially produce the “discovery” of a significant effect impacting our conclusions on the “populations” being compared. Usually, H0 states that the tested levels of a given factor come from the same population, meaning that no level is significantly different from the other. Given that H0 is true, testing many levels against each other increases the chances of erroneous “discovery” solely due to sampling! Therefore raising the accepted risk of rejecting H0 (when it is true) from α to αf>α.

R functions (summary, anova, contrasts, car::Anova(), emmeans(), pairs(), mvcontrast(), )

## Module 5 - Resampling and Simulation Methods for MLM

***(Term Project focus)***

* Confidence intervals (CI) provide an alternative inferential procedure to hypothesis testing.
* They do not require the statement of a null hypothesis H0H0, and provide information about the possible range of values of a parameter (which often translates in an interpretable quantity).
* Inferential procedures for MLM, in particular, such as tests and CIs, are usually valid under some conditions.
* One (often) important condition is that the sample is sufficiently large, for asymptotic theory to be accurate enough.
* One however does not know, in general, what large enough means.
* An alternative inferential procedure to asymptotic theory is given by bootstrap methods.
* They simulate the empirical distribution of the estimator θ^ or the test statistic T, by generating samples (similar to the observed one), using an estimated model.
* The estimated model can be the assumed model with parameters replaced by a consistent estimator, or the empirical distribution generated by draws without replacement from the observed sample.
* With MLM, the latter procedure is difficult to implement when there are covariates which are factors, since conditions such as balanced groups should be respected.
* Anyway, before basing a decision given a sample of data, such as the outcome of a CI or a test, one should check, using MC methods, if the chosen procedure is accurate enough.
* For CIs, the accuracy is measured by comparing the theoretical confidence level to the empirical one using MC simulations.
* For tests, the accuracy is measured by comparing the theoretical significance level with the empirical one, using MC simulations.

*Supplementary material* - (A Practical Guide to Monte Carlo Simulation for MLM)

Perform *Monte-Carlo (MC) simulations*.  
Using the overall *architecture* of a model, i.e. the full model specification, defined by its design matrices and model assumptions, we show how to generate random vectors of responses for a chosen set of parameter values.

These random response vectors, together with the design matrices, represent simulated samples for which different statistics (estimates, test statistics, CI) can be computed, to build up a *sampling distribution* for the latter. Then, for a given model, parameters values and sample size, this allows to e.g.

* assess the properties of an estimator for the chosen parameter (unbiasedness, convergence),
* compare estimators (unbiasedness, efficiency, mean square error),
* assess if a method used to perform a test has the correct (finite sample) significance level α,
* assess if a method used to compute CI has the correct (finite sample) coverage at the confidence level 1−α

## Generating MLM samples

Similarly to the LM, the MLM can be expressed as follows:

y=Xβ+Zu+εy=Xβ+Zu+ε

where

* y, X and ε are defined as in the previous section,
* Z is the (N×q) random effect design matrix (defined by staking (in columns) the design matrices of the rr individual random effect matrices, Z1, …, Zr, of size (N×qr) with q=∑ri=1qrq=∑i=1rqr
* u is the (q×1) random effect vector with u∼N(0q,Ψ) where Ψ denotes the random effect covariance matrix.

In this section, we learn how to simulate response vectors for MLM given

* the chosen sample size and (X and Z) design matrices,
* chosen parameter vector values, i.e, values for θ=[βT,σ21,...,σ2r,σ2ε]T.

Hence, compared to a LM, the additional required elements are

* the design matrix Z,
* the distribution of the random effects uu including the random effect variances (and covariances in case of correlated random effects)

## Module 6 - Prediction and Model Validation

* A statistical model, such as an MLM, needs to be validated before any conclusion is taken on the basis of the data.
* There might be several types of departures from the model's assumptions, such as nonlinearity in the (conditional) expectation (i.e. the linear predictor), outlying observations, non normality of the random effects and the residual error, non independence between random effects and the residual error, non constant variance of the resiudal error, etc.
* Although a descriptive analysis can already help in setting up the most precise form for the model, a posthoc analysis is very much necessary.
* For that, one can study the behavior of estimated quantities, which for MLM, are the estimated random effects, the fitted responses and the residuals (estimated residual error).
* The analysis of these quantities is performed graphically, in order to detect potential departures from the assumed model.

*Supplementary material* - (Model Validation)

Model validation for a MLM can be performed using residuals and random effects estimates. In `R’, these can be obtained in several ways. Useful functions include:

* **random.effects()** estimates the random effect associated with every experimental unit
* **predict()** allows to estimate both the population and the subject predictions or fitted values.
* **fitted()** do the same.

The aim of the model check it to assess if:

* the normality assumption is reasonable for random effects and residuals (normal qq-plots)
* the assumption of homoscedasticity seems respected (boxplot of the random effect estimates and residuals versus fixed variables, plot of the residuals versus fitted values)
* the assumption of linearity of the relationship between fixed predictors and the response (plot of the residuals versus fixed variables)
* the model fit may be influenced by outliers