# NRES 798 — Statistical Methods for Ecologists
# Hierarchical and mixed effects models

Oscar García

April 5, 2013

## Contents



## 1 Introduction

Mixed effects models is a fashionable and controversial topic. Most presentations and applications are somewhat obscure in relation to the underlying reasoning, assumptions, and motivation. A great deal of black magic and "expert opinion" tends to be involved. This is an attempt at explaining as clearly and precisely as I can my understanding of the subject. It is possible that I might be oversimplifying and missing out on something very profound, although I doubt it. It is useful to study this together with the material in Lab 12.

These models are the most important tools used in so-called Multilevel Analysis, which deals mainly with hierarchical models. The three concepts are interrelated and sometimes used as synonyms. They have in common the idea of having several types of unit of analysis, not just one as in most of the rest of statistical inference. Often these are nested into a number of hierarchical levels, such as students within classrooms within schools, employees within departments within firms, or successive measurements within trees within stands within forests. However, levels in multilevel analysis may also be *crossed* rather than *nested*, they may belong to a relationship network more general than a hierarchical tree. And mixed effects models can deal with levels (or effects) that are not in a hierarchy. And as seen below, multilevel and hierarchical data can be modelled and analyzed without mixed models. But the hierarchical case with levels nested within other levels is the most common, and the examples will focus on that.

A good entry point to the literature is the article by Snijders on *Multilevel Analysis* in *The SAGE Encyclopedia of Social Science Research Methods*, 2003.

## 2 Multilevel models, fixed effects

We use a simple example, see Lab 12 for computational and other details. On 14 loblolly pine trees, heights have been measured at 5-year intervals. We use as the response the 5-year height increments, and as a predictor the height at the start of the interval; there are 4 growth measurements for each tree[1].

Pooling all the data, the following simple linear regression model appears reasonable:

$$Y_{ij} = \alpha + \beta x_{ij} + \varepsilon_{ij} \ , \tag{1}$$

where $Y_{ij}$ is the $j$-th growth increment for tree $i$, $x_{ij}$ is the corresponding height, and $\alpha$ and $\beta$ are regression parameters. As usual, the $\varepsilon_{ij}$ are assumed to be independent normal random variables with mean 0 and variance $\sigma^2$. Least-squares gives the estimates $\hat{\alpha} = 18.95$ and $\hat{\beta} = -0.1985$, with RSE $= \hat{\sigma} = 0.9243$.

This line represents an "average" relationship between growth and height. But the trees differ among them in genetic makeup and local growing conditions. One may be interested in how the relationship might vary among

[1] It is well known that, within one tree, size observations over time tend to be autocorrelated. Because size is the accumulation of growth increments that vary due to weather and other factors. However, weather, etc., is independent or only weakly correlated between one time interval and the next, so that size increments can be expected to be approximately independent. Unless there is a large size measurement error, which would introduce correlation because consecutive increments share one size measurement. The increment decreasing linearly with size corresponds to the so-called *Mitscherlich* or *monomolecular* growth model.

trees. We can try with a common slope $\beta$, but letting each tree to have its own different intercept $\alpha_i$:

$$Y_{ij} = \alpha_i + \beta x_{ij} + \varepsilon_{ij} \ .$$

Or in a more standard linear model notation,

$$Y_{ij} = \alpha + \Delta\alpha_i + \beta x_{ij} + \varepsilon_{ij} \ , \tag{2}$$

where $\alpha$ is the average $\alpha_i$ among all the trees, and $\Delta\alpha_i = \alpha_i - \alpha$. The estimates for the $\Delta\alpha_i$ should be constrained to add to 0, the typical *sum* contrast to avoid redundancy.

Now, in addition to height, there is a categorical variable identifying the trees. In the Wilkins-Rogers computer notation this is something like `growth ~ tree + height`. The estimates obtained are $\hat{\alpha} = 19.08$, $\hat{\beta} = -0.2023$, RSE $= \hat{\sigma} = 0.7027$, plus 14 $\widehat{\Delta\alpha_i}$ in which we are usually not interested. An $F$-test indicates that the variable-intercepts model (2) is "significantly better" than (1).

Similarly, one can fit a model with a common intercept and variable slopes, or one with both $\alpha$ and $\beta$ varying among trees. See Lab 12. They were not significantly better than the variable intercept model.

This is a two-level hierarchical linear model[2], growth measurements are nested within trees. With balanced data, as in here, the hypothesis tests are an instance of analysis of covariance (ANCOVA), and there is a similar example in Chapter 10 of the textbook. The general linear model approach and the $F$-tests are valid for unbalanced data too.

In general, in a multilevel model the observations belong to a number of groups or sub-populations. These are nested in two or more levels in hierarchical models. In a fixed effects formulation one has a linear or nonlinear regression model

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) \ , \tag{3}$$

where some of the parameters in the parameter vector $\boldsymbol{\beta}$ are specific (local) to one or more groups, while others are common to all the groups (global). The predictors may belong to different levels.

Reparametrization is commonly ised in building the model. With two levels, one may start with an ordinary regression model like (3) with $\boldsymbol{\beta}$ global. Then,

[2] Although this name is usually reserved for the mixed effects version.

$\boldsymbol{\beta}$ is replaced by a function of new global and local parameters. The function may contain predictors. In the example we simply substituted $\alpha + \Delta\alpha_i$ for $\alpha$. The process can be repeated for more levels.

# 3 Mixed effects

Continuing with the example, in model (2) there are many parameters to be estimated: $\alpha, \beta, \sigma$, and 13 independent $\Delta\alpha$'s. Often one is not interested in the values of the locals, in statistical jargon they are *nuisance parameters.*

A different way of modelling this is to assume that the trees are a simple random sample from a large population of trees, where the $\alpha_i$'s or $\Delta\alpha_i$'s have a normal distribution. Instead of $\Delta\alpha_i$ being an unknown parameter, it is assumed to be a normal random variable $\varepsilon_i$ with mean 0 $\alpha$ and an unknown variance $\eta^2$. This is called a *random effect.* The model becomes

$$Y_{ij} = \alpha + \varepsilon_i + \beta x_{ij} + \varepsilon_{ij} \ . \tag{4}$$

Now there are only 4 parameters to be estimated: $\alpha, \beta, \sigma$, and $\eta$.

This is a *linear mixed effects model.* Ordinary regression models have only one random variable $\varepsilon$, these models have more than one. The parameter estimates in $R$, using `lme` from package `nlme`, were $\hat{\alpha} = 19.05, \hat{\beta} = -0.2014, \hat{\sigma} = 0.7029$, and $\hat{\eta} = 0.6134$ (see Lab 12). Very similar to those from (2).

In general, one has a model like (3), but where instead of local parameters one has normal random variables (random effects). The random effects within a group or level may be correlated, they are assumed to have a multivariate normal distribution with a covariance matrix whose elements may be partially or totally unknown.

There are methods for "estimating" values of the random effects for the relevant groups, similar to the local parameter estimates in the fixed effects approach. Of course, estimating a random variable does not make sense from the classical statistics point of view, so a somewhat *ad hoc* reasoning inspired in Bayesian ideas is used.

Although the mixed effects estimation algorithms are considerably more complicated than those for linear or nonlinear regression, there can be computational savings over to estimating a fixed effects model if the number of local parameters is large. A large number of parameters might also reduce precision in the (global) parameters of interest. ANOVA fans might say that many “degrees of freedom are lost” estimating parameters about which we do not care anyway. On the other hand, mixed effects models introduce strong additional assumptions that may or may not be reasonable. In particular, a mixed effects model would seem to be inappropriate if the selection of analysis units is design-based, and not a simple random sample from the target population.