# Specifying linear mixed models in lme4

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

July 25, 2008

**Abstract**

A linear mixed model to be fit by `lmer` is specified by a formula. The fixed-effects terms in this formula are interpreted as they would be in the `lm` or `glm` functions. In this vignette we describe the interpretation of the random effects terms in these formulas and provide several examples of fitted models and their formulas.

## 1 Introduction

We begin with examples of data sets to which we will fit some linear mixed models. Once we have some examples to bear in mind we will describe the mathematical formulation of the model and how the formula describing the model in `lmer` is interpreted.

## 2 Examples of data sets and models

**?** is a classic reference on the use of statistics in the chemical industry. The first edition was published in 1947. Although we refer to the chapters and pages in the fourth edition, published in 1972, the discussion of these data sets and models to be fit to them does go back to 1947 and earlier.

### 2.1 The Dyestuff example

The `Dyestuff` data in the `lme4` package, shown in Figure 1, are described in

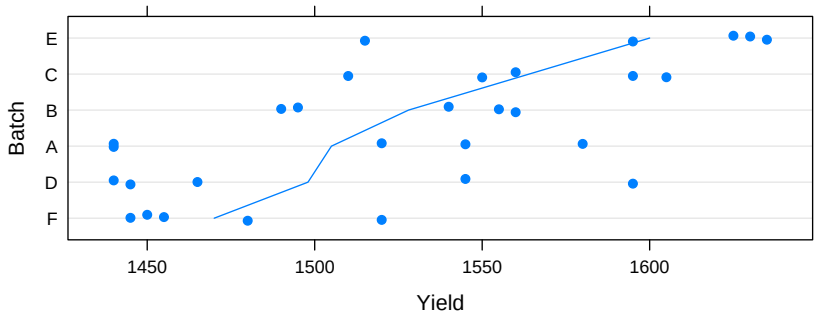


Figure 1: Dotplot of the data from the dyestuff example described in Davies, 1972. The six batches of the intermediate product determine the row. Within each row the yields for the five preparations from that batch are shown. The points have been jittered slightly on the vertical axis to prevent overplotting. Because there is no indication of a preferred ordering for the batches (such as a time ordering) we reorder the batches according to increasing mean yield. The line joins the mean yields of the batches.

Example 6.1 of (**?**, p. 130) as coming from

> an investigation to find out how much the variation from batch to batch in the quality of an intermediate product (H-acid) contributes to the variation of the yield of a dyestuff (Napthalene Black 12B) made from it. In the experiment six samples of the intermediate, representing different batches of works manufacture, were obtained, and five preparations of dyestuff were made in the laboratory from each sample. The equivalent yield of each preparation as grams of standard color was determined by dye-trial, ...

Note that the purpose of the experiment is to characterize the variation in the quality of the product that can be attributed to the batch to batch variation of the intermediate product. The yield is the response and the batch is the covariate. Batch is a *categorical* covariate, in the sense that the information about the batch is simply whether the sample was created from the first batch or the second batch or so on. We can take any of the 30 observed yields and categorize it as having been prepared from one of the six batches of intermediate.

The *factor* data type in R provides a representation for such categorical covariates. In contrast, the yield is a numerical response measured on a physically meaningful scale (grams) and we represent that as a numerical value.

```
> str(Dyestuff)
'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ Yield: num  1545 1440 1440 1520 1580 ...
```

We say that there are six *levels* of the `Batch` covariate. When incorporating categorical covariates like `Batch` in a linear statistical model we obtain numerical values for the *effects* of the different levels. We can see from Figure 1 that batch F had the lowest mean yield and batch E had the highest mean yield so in our model we expect to see a low effect for batch F and a high effect for batch F.

We distinguish between the effects of factors with a fixed and reproducible set of levels, such as the sex of a participant in a experiment on human or animal subjects, and those of factors for which the observed set of levels can change throughout the experiment or study. These are, not surprisingly, called *fixed effects* and *random effects*, respectively. In a way these terms are misleading because it is the set of levels of the factor associated with the effects that we determine to be fixed or random, not the effects per se. Nevertheless, these terms are widely used and hence we adopt them.

We employ fixed effects and random effects terms in a statistical model for different purposes. Because a fixed-effects term is associated with a fixed set of levels for a factor, we want to estimate the effect of those particular levels. Often we also want to contrast the effects of particular levels of the factor. In a clinical trial, for example, some patients may receive Drug A, some may receive Drug B and some may receive a placebo. Typically the purpose of the trial is to contrast the effects of specific drugs or to compare the effects on patients a particular drug versus a placebo.

In a random-effects term the effects of particular levels of the factor are not of as much interest to us as is the amount of variation in the response that can be attributed to the different levels of the factor. This is exactly the situation in this dyestuff experiment. The batches that were examined are but a sample of the batches that could be or have been produced. For the purposes of predicting future yields we are not interested in the effects of past batches, which may already have been used up, as we are in the effects of future, as yet unobserved, batches. We can't know exactly what their levels

may be but we can characterize the batch to batch variability that we have seen and base our predictions on that.

A *mixed-effects* model is a statistical model that incorporates fixed-effects parameters and random effects. As we shall see in §3.1, the mathematical formulation of random effects that we use requires that there always be at least one fixed effect in the model. That is, in our formulation any model that incorporates random effects is a mixed-effects model.

We can fit a linear mixed-effects model to the response `Yield` in the `Dyestuff` data incorporating random effects for the `Batch` factor, save the fitted model as `Dm1` then summarize it with

```
> summary(Dm1 <- lmer(Yield ~ 1 + (1 | Batch), Dyestuff))
Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
 AIC BIC logLik deviance REMLdev
 326 330   -160      327     320
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1764     42.0
 Residual             2451     49.5
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   1527.5       19.4    78.8
```

## 2.2 The Dyestuff model formula

The `lmer` function (the name, pronounced like "Elmer", is an acronym for Linear Mixed-Effects in R) follows the convention of most model fitting functions in `R` that the first two arguments are `formula`, a formula describing the model, and `data`, the optional name of a data frame in which the formula can be evaluated.

The formula for model `Dm1`

```
Yield ~ 1 + (1 | Batch)
```

can be read as "`Yield` is modeled by a constant plus a constant given `Batch`". This formula consists of two *terms*, `1` and `(1|Batch)`. In general, terms in the model are separated by plus signs (`+`). A term incorporating the vertical bar character (`|`) is a random-effects term. A term without the vertical bar is a fixed-effects term.

There is a single fixed-effects term, `1`, in this formula. A model matrix, which we will write as $\boldsymbol{X}$, is created by evaluating all the fixed-effects terms

in the formula using the data frame and stored in the slot named `X` in the fitted model. In this case $\boldsymbol{X}$ is a trivial model matrix with 30 rows and one column. All of the elements of $\boldsymbol{X}$ are unity. The first three rows of this matrix are

```
> head(Dm1@X, n = 3)
     [,1]
[1,]    1
[2,]    1
[3,]    1
```

The single random-effects term in this formula is `(1|Batch)`. In a random-effects term the expression on the right hand side of the vertical bar must evaluate to a factor. Typically it is simply the name of a factor, like `Batch` here, but more general expressions are possible. The expression on the left hand side in the data frame as a linear model formula, producing a model matrix. In this case the model matrix from the left hand side is the same as $\boldsymbol{X}$. It has one column and 30 rows.

# 3 Mathematical formulation of the model

A linear mixed-effects model (LMM) is statistical model similar to the conventional linear model (also called a linear regression model). In both types of models we consider a set of observed responses, which we shall write as the $n$-dimensional vector $\boldsymbol{y}$, and associated values of other variables, which we shall call *covariates*. These models are linear in the sense that we express the effect of the covariates in terms of *model matrices*.

For example, a linear model is frequently written

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{1}$$

where $\boldsymbol{\beta}$ is a $p$-dimensional parameter vector, $\boldsymbol{X}$ is an $n \times p$ model matrix derived from the model formula and the observed values of the covariates, and $\boldsymbol{\epsilon}$ is the random noise, or unexplained variation, in the observations. As indicated in (1) we typically begin with the assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. That is, the noise is assumed to have a multivariate normal (or "Gaussian") distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the identity matrix of size $n$.

Writing a linear model as (1) blurs the distinction between the random variable $\boldsymbol{\mathcal{Y}}$ and its observed value $\boldsymbol{y}$. Because this distinction is important in describing LMMs we will rewrite the linear model (1) as

$$\boldsymbol{\mathcal{Y}} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right) \tag{2}$$

## 3.1 Probability distribution formulation of LMMs

A linear mixed-effects model incorporates fixed-effects parameters and random effects. Technically, the random effects are unobserved random variables, which we will write as the random vector $\boldsymbol{\mathcal{B}}$, while the fixed-effects parameters are indeed parameters. We will write them as $\boldsymbol{\beta}$. The random variable representing the response is $\boldsymbol{\mathcal{Y}}$ with observed value $\boldsymbol{y}$. For a linear mixed model, the conditional distribution of $\boldsymbol{\mathcal{Y}}$, given $\boldsymbol{\mathcal{B}} = \boldsymbol{b}$, is a multivariate normal (or "Gaussian") distribution with (conditional) mean $\boldsymbol{X\beta} + \boldsymbol{Zb}$ and variance-covariance matrix $\sigma^2\boldsymbol{I}_n$ where $n$ is the number of observations (the dimension of $\boldsymbol{y}$) and the notation $\boldsymbol{I}_n$ indicates the $n \times n$ identity matrix of size $n$. The fixed-effects parameter vector $\boldsymbol{\beta}$ is of dimension $p$ and its model matrix $\boldsymbol{X}$ is $n \times p$. The random effects are of dimension $q$ and their model matrix $\boldsymbol{Z}$ is $n \times q$.

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{X\beta} + \boldsymbol{Zb}, \sigma^2\boldsymbol{I}_n\right) \tag{3}$$

The distribution of $\boldsymbol{\mathcal{B}}$ is also assumed to be multivariate normal, this time with mean $\mathbf{0}$ and a $q \times q$ symmetric variance-covariance matrix that we will write as $\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$ where $\sigma^2$ is the same parameter used in the variance-covariance of $\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}$ and $\boldsymbol{\Sigma}$ is a $q \times q$ *relative variance matrix* for the random effects. The notation $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ indicates that $\boldsymbol{\Sigma}$ depends on a parameter vector $\boldsymbol{\theta}$. Typically the dimension of $\boldsymbol{\theta}$ is much, much smaller than $q$, the size of $\boldsymbol{\Sigma}$.

The model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, the form of $\boldsymbol{\Sigma}$ and how $\boldsymbol{\Sigma}$ depends on $\boldsymbol{\theta}$ are all specified by the formula which is the first argument to `lmer`.