Linear statistical models,
with an introduction to mixed effects models

Christoph Scherber, September 2008

Recall that a “classical” linear model can be written as:

response variable = explanatory variable(s) + Error

For example, if we are performing a simple linear regression, we generally use a model of the form

(1) $y = a + bx$

where a is the intercept and b is the slope. In many statistics textbooks, the parameters a and b will be labeled $ß_0$ and $ß_1$ by convention. And because every observation usually doesn´t lie perfectly on the regression line, it usually is accompanied by an error component, $\varepsilon_{i.}$ The i stands for the i´th observation:

(2) $y_i = ß_0 + ß_1 x_i + \varepsilon_i$

Now how can be generalize this expression for all linear models? We can, for example, write that every observation y is a function of a systematic component, $\mu(x; ß)$, and an error component, ε. **ß** is a vector of parameters, and x is a vector of explanatory variables. This is why we write both in bold:

(3) $Y(\mathbf{x}, \varepsilon) = \mu(\mathbf{x}; \mathbf{ß}) + \varepsilon$

Here, all the **x**´s are somewhat controllable, non-random variables that are usually called explanatory variables (or independent variables). Note that the systematic component, $\mu(\mathbf{x};\mathbf{ß})$, and the errors, ε, are assumed to be additive.

Before we dive a bit deeper into linear models, let us think of some examples:

Y could be the yield (biomass) of a piece of arable land
x could be fertilization, watering (irrigation) or seeds added
ε could be random variation due to weather or soil conditions

Likewise, for an observation on growth of grasshoppers in a field,

Y could be the weight of a single grasshopper (from a population of 20)
x could be the age or whether it was a female or a male,
ε could be random variation due to its individual health status

You see the point. All the Y´s are the things we observe, x are things that we can directly influence as experimenters, and ε are errors that we cannot directly influence or know.

Now why do we want to express our linear models using the systematic component $\mu(\mathbf{x};\ \mathbf{ß})$? The answer is that we can express many different kinds of models using this very general approach.

For example, if we wanted to find an expression for a non-linear statistical model, we might want to use

$$(4)\ \mu(\mathbf{x};\ \mathbf{ß}) = ß_0 x/(ß_1+x),$$

which would correspond to a Michaelis-Menten non-linear saturating curve.

Suppose now that we wanted to generalize all our linear models using a similar approach. Then we can express $\mu(\mathbf{x};\ \mathbf{ß})$ as a linear combination of known functions $g_j(\mathbf{x})$ with unknown coefficients $ß_j$:

$$(5)\ \mu(\mathbf{x};\ \mathbf{ß}) = \sum_{j=0}^{m} ß_j g_j(x)$$

We can even make this expression simpler by defining x_j as the result of $g_j(x)$:

$$(6)\ x_j=g_j(x)$$

The complicate equation from above now becomes

$$(7)\ \mu(\mathbf{x};\ \mathbf{ß}) = \sum_{\mathbf{j=0}}^{\mathbf{m}} ß_j \mathbf{x_j}$$

And, if we also extract ß0 from the equation, we get

$$(8)\ \mu(\mathbf{x};\ \mathbf{ß}) = ß_0 + \sum_{\mathbf{j=1}}^{\mathbf{m}} ß_j \mathbf{x_j}$$

What we have done so far is formulate expressions for our systematic component, $\mu(\mathbf{x};\ \mathbf{ß})$. Remember that each linear model consists not only of this systematic component, but also of an error component:

$$(9)\ y\ (\mathbf{x},\ \varepsilon)= \mu(\mathbf{x};\ \mathbf{ß}) + \varepsilon$$

If we bring all this together, we can write:

$$(10)\ y\ (\mathbf{x},\ \varepsilon)= \mu(\mathbf{x};\ \mathbf{ß}) + \varepsilon$$

$$(11)\ y\ (\mathbf{x},\ \varepsilon)= \sum_{\mathbf{j=0}}^{\mathbf{m}} ß_j \mathbf{x_j} + \varepsilon$$

Note that all the bold letters indicate we are dealing with matrices and vectors. Suppose we want to get rid of this:

$$(12)\ y_i = \mu_i + \varepsilon_i = \sum_{\mathbf{j=0}}^{\mathbf{m}} \mathbf{x_{ij}} \mathbf{ß_j} + \varepsilon_i$$

This equation is exactly the same as equation (11), except for that we do not express it in matrix form any more.

Now here are several equations that are all exactly equivalent:

$$(13)\ y = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

$$(14)\ y = \mathbf{X}\mathbf{ß} + \boldsymbol{\varepsilon}$$

$$(15)\ y = \sum_{j=0}^{m} ß_j \mathbf{x}_j \ + \varepsilon$$

$\mathbf{y} = (y_i;\ldots;y_n)^T$
$\boldsymbol{\mu} = (\mu_1;\ldots;\mu_n)^T$
$\boldsymbol{\varepsilon} = (\varepsilon_1;\ldots;\varepsilon_n)^T$
$\mathbf{X} = (\mathbf{x}_0;\ldots;\mathbf{x}_m)$

Here, the following abbreviations have been used:

$\mathbf{y} = (y_i;\ldots;y_n)^T$ is the vector of the **response variable**
$\boldsymbol{\mu} = (\mu_1;\ldots;\mu_n)^T$ is the vector of the **systematic component**
$\boldsymbol{\varepsilon} = (\varepsilon_1;\ldots;\varepsilon_n)^T$ is the vector of the **error component**
$\mathbf{X} = (x_0;\ldots;x_m)$ is the **design matrix**

The **design matrix** is something that may be unfamiliar to you. It consists of rows and columns (like any matrix). The rows in this matrix are the “replicates” of your experiment – there are as many rows as there are individual data points. The design matrix has as many columns as there are explanatory variables. The design matrix contains the values each explanatory variable assumes. In case of the intercept, these values will always be 1, because the design matrix is later multiplied by the vector of coefficients, ß.

Suppose we are measuring yield in 4 plants, and the amount of fertilizer added is used as an explanatory variable. Then the design matrix will look like this:

Intercept	Fertilizer added
1	10
1	20
1	30
1	40

Likewise, if we are dealing with an ANOVA model, we might just want to have plants that are either fertilized or not. Then, the design matrix becomes:

Intercept	Fertilized
1	0
1	0
1	1
1	1

Mixed-effects models

What is the difference between a linear model and a linear mixed-effects model? Well, it is easiest to find this out by looking at a mixed-effects model and compare it to what we already know:

Linear model: $y = \mathbf{X\text{ß}} + \boldsymbol{\varepsilon}$
Mixed effects model: $y = \mathbf{X\text{ß}} + \mathbf{Zb} + \boldsymbol{\varepsilon}$

Here, there is a fixed effect ß for every element in the fixed-effects matrix X. In addition, there is a random effects component b. By convention, **random effects** are always denoted by **Latin** characters, while **fixed effects** get **Greek** characters. Z is an identity matrix just consisting of one column of 1´s.

The b´s are all assumed to be normally distributed with mean 0 and one standard deviation.

More details will be added soon.