Linear Mixed Models

Paul Schmidt

2023-12-12

What are random effects and linear mixed models?

Table of contents

1 What even is a random effect?

Typically, learning about random effects in statistical methods comes after grasping simpler regressions and ANOVAs. As a result, when someone mentions "an effect" in statistics, it often refers to a fixed effect, making random effects somewhat special and explicitly labeled as such. While fixed effects models date back to the early 19th century (Legendre, 1805; Gauss, 1809), the understanding and application of random effects in statistical models are relatively newer concepts.

It could be noted that the error term (e) in all models, including those with only fixed effects, is a random variable, representing a stochastic component. Thus, strictly speaking, nearly all models could be considered mixed models, but the error term plays a unique role and is not counted as a random effect.

When random effects are explained, they are often described as representing random samples from a population or factors whose levels, unlike fixed effects, are randomly selected from a larger set.

Another common point of discussion is the minimum number of factor levels required before considering a factor as random. Recommendations usually vary between 5 and 12 levels.

1.1 When is a random effect?

Here are some rules of thumb that call for a random effect:

- When factor levels are randomly selected from a larger population.
  - Examples
    - Selecting a few locations from a larger target region.
- When there are approximately 8 or more levels of the factor.
- When it represents an additional randomization unit in an experiment's design.
  - Examples
    - Mainplots in split-plot designs
    - Subplots in split-split-plot designs
  - Relevant chapters
    - Split-plot Design
- When it involves incomplete blocks in an experiment.
  - Relevant chapters
    - Alpha Design
    - Augmented Design
- When it represents sub-samples of experimental units.
  - Examples
    - Multiple plants sampled within the same plot (at the same time)
    - Multiple leaves sampled from the same plant (at the same time)
    - Multiple measurements sampled from the same animal (at the same time)
  - Relevant chapters
    - Sub-sampling
- When it is crossed with another random effect.
  - Examples
    - If the main effect of 'Location' is random, then the 'Year:Location' interaction effect is automatically random as well.

- When assuming correlations or specific variance structures between levels.
  - Examples
    - Between genotypes due to genetic similarity
    - Between repeated measures on the same experimental unit due to temporal proximity (=over time)
    - Between samples on the same field due to spatial proximity
  - Relevant chapters
    - Repeated Measures

2 General Representation

simple linear model

$$y = X\beta + e$$

linear mixed model

$$y = X\beta + Zu + e$$

2.1 Immediate observations

In this formula, y is a vector of observations (e.g., yield values). X is the design matrix for fixed effects, and β is the vector of these fixed effects. Analogously, Z is the design matrix for random effects, and u is the vector of random effects. Therefore, the difference between a linear model and a linear mixed model is Zu.

Before a model is fitted, β and u are unknown, meaning we don't yet have numerical values for these effects – they need to be estimated. However, y, X and Z are known quantities; they contain the measured values and, in simple terms, the treatments or conditions that these measurements have undergone. The design matrices X and Z connect the effects to the observational values.

Lastly, we have the random vector of errors e. It's important to recognize that the errors in the model are neither known before the model fitting, nor are they explicitly estimated as part of the model's output. Rather, they inherently represent the residual portion of the data that the fitted model fails to explain. However, these errors are not entirely arbitrary; they are constrained by certain conditions. Specifically, they collectively sum to zero and are typically assumed to have a uniform variance across observations. This implies that while individual errors may vary, they do so within a structured framework set by the model.

It's crucial to understand that the length of y and e, and the number of rows in X and Z, correspond to the number of observations in the dataset. The length of β depends on how many fixed effects are to be fitted in the model (at least 1, i.e. μ) and matches the number of columns in X. The same logic applies to u for random effects and Z.

2.1.1 An example

Let's consider a simple linear model $y = X\beta + e$ using the following dataset:

```
data <- data.frame(
  var = factor(c("A", "A", "B", "B", "C", "C")),
  yield = c(3.2, 3.6, 2.8, 2.9, 4.1, 4.0)
)
```

```
data
```

```
  var yield
1   A   3.2
2   A   3.6
3   B   2.8
4   B   2.9
5   C   4.1
6   C   4.0
```

When fitting the following model, this happens:

```
mod <- lm(yield ~ var, data = data)
```

$$\begin{bmatrix} 3.2 \\ 3.6 \\ 2.8 \\ 2.9 \\ 4.1 \\ 4.0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ var_B \\ var_C \end{bmatrix}$$

After the model is fitted, we find this:

```
coefficients(mod) # show estimated beta
```

```
(Intercept)        varB        varC
       3.40       -0.55        0.65
```

```
resid(mod) # show residuals
```

```
    1     2     3     4     5     6
-0.20  0.20 -0.05  0.05  0.05 -0.05
```

$$\begin{bmatrix} 3.2 \\ 3.6 \\ 2.8 \\ 2.9 \\ 4.1 \\ 4.0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3.40 \\ -0.55 \\ 0.65 \end{bmatrix} + \begin{bmatrix} -0.20 \\ 0.20 \\ -0.05 \\ 0.05 \\ 0.05 \\ -0.05 \end{bmatrix}$$

2.2 Less obvious details

Simply knowing the components mentioned above doesn't fully describe a (mixed) model. Typically, one should also mention the distributions and covariance structures:

$u \sim MVN(0, G)$

$e \sim MVN(0, R)$

$y \sim MVN(X\beta, V)$

Unless stated otherwise, a multivariate normal distribution is often assumed, denoted here as $MVN()$. If the distribution were not normal, we would be discussing generalized mixed models.

The first entry within the parentheses indicates the expected value (mean vector). Both random effects u and errors e have an expected value of 0. This makes sense as they represent only random deviations from the fixed part of the model. Thus, individual errors may be positive or negative, but on average, their expected value is 0. The expected value for y, our observations, is naturally not 0 but depends on the data and is expressed by the fixed effects estimated from the data.

For instance, consider a model with μ as the only fixed effect. Then $\beta = \mu$ and X is a vector filled with 1s, its length corresponding to the number of observations. In such a simplified case, the entire fixed part of the model is reduced to an overall mean μ, and the expected value for our observations y would be this overall mean.

3 ML & REML

Unlike simple linear models, mixed models (and generalized models) cannot rely solely on the Least Squares Method (LSM) (Legendre, 1805) for estimating all their parameters. Instead, methods like the Maximum Likelihood (ML) approach (Edgeworth, 1908) are necessary. As the name suggests, ML estimates parameters by maximizing a likelihood function so that the

parameter distribution best matches the observed data. The estimates are the most probable or maximum likely. However, it is essential to specify the distribution for which this maximization is performed. Interestingly, if a simple linear model (like a linear regression assuming normal distribution) is estimated using ML, the parameter estimates (β) will be identical to those obtained via LSM.

One advantage of the ML method is its flexibility. It can be used for models that do not assume a normally distributed error term. Instead, it can accommodate distributions from the Exponential Family, which includes, besides the normal distribution, the binomial, Poisson, gamma, and inverse Gaussian distributions.

However, a drawback of the ML method, especially in mixed models (even under normal distribution), is its tendency to underestimate the true variance components, particularly in small samples.

On the other hand, the Restricted Maximum Likelihood (REML) method (Bartlett, 1937) offers an unbiased estimation of variance components. In mixed models, REML is used specifically to estimate V, not β. The REML estimation of V is then utilized to perform an ML estimation of β. This implies that the variance components estimated via REML are independent of the fixed effects in the model.

One limitation of REML, compared to ML, is that the model fits from different models for the same dataset are only comparably valid as long as they contain the same fixed effects. In other words, the goodness of fit of various REML models should not be compared if they have different fixed effects. This limitation restricts the scope of REML in comparing model fits across different models with varying fixed effects.

TODO