The Aguirre Lab
DePaul University

The Warner Lab
Auburn University

Creating Linear Mixed Models in R-Studio

Linear mixed models (LMMs) are a type of statistical model used to analyze data with both fixed and random effects. In an LMM, the response variable is modeled as a function of **fixed effects** (variables that have a known and measurable effect on the response variable) and **random effects** (variables that are not of primary interest but are included in the model to account for the variance in the response variable that cannot be explained by the fixed effects).

The LMM is a powerful tool for analyzing data in many areas of biology, including evolutionary biology, conservation biology, and ecology. For example, LMMs can be used to analyze how environmental factors influence the evolution of plant phenotypes (Ghalambor et al., 2007), to model the population dynamics of threatened or endangered species (Coulson et al., 2011), or to study the effects of climate change on ecological communities (Bell et al., 2013).

To create an LMM in R, you will need to install the 'lme4' package. You can install the package using the following lines of code:

Installation Commands

Macintosh	**install.packages("lme4", type = "source")**
Windows	**install.packages("lme4")**
Linux	**sudo apt-get install r-cran-lme4**

Once you have installed the package, you can set up an LMM using the example code in **Box 1**.

Box 1: Example Code for Running a Linear Mixed Model

```
# Load the 'lme4' package
library(lme4)

# Set up the linear mixed model
model <- lmer(Response ~ FixedEffect1 + FixedEffect2 + (1|RandomEffect1) +
(1|RandomEffect2), data = dataset, REML = TRUE) *or FALSE*

# Print the summary of the model
summary(model)
```

'**Response**' is the response variable.

'**FixedEffect1**' and '**FixedEffect2**' are the fixed effects.

'**RandomEffect1**' and '**RandomEffect2**' are the random effects.

The **'data**' argument specifies the dataset used for the analysis.

'**REML**' refers to using restricted maximum likelihood principles.

In linear mixed models, both fixed and random effects are used to account for variation in the data. Fixed effects refer to variables that are predetermined and known, and they are typically used to model the average effect of a treatment or an intervention. Fixed effects can be represented by variables that are coded as dummy variables or indicator variables (Gelman & Hill, 2007). These variables are predetermined, and their values do not vary across observations. Fixed effects can be used to model the impact of a treatment or intervention on the outcome of interest, or to adjust for potential confounding variables.

Random effects, on the other hand, refer to variables that are not predetermined and can vary from one observation to another. They are used to model individual differences between observations. Random effects are represented by variables that have a distribution of possible values. They are used to account for variation that is not explained by the fixed effects. For example, in a study of student test scores, random effects might be used to account for differences between classrooms or between schools (Pinheiro & Bates, 2000).

Overall, the use of fixed and random effects in linear mixed models allows for a more comprehensive understanding of the factors that influence the outcome of interest and can help to improve the accuracy of statistical inference.

Now, let's work with a trial data set.

Data.gov can be a valuable resource for bioinformaticians who are looking for publicly available datasets to use in their research. Some of the datasets available on Data.gov may be relevant to bioinformatics, such as those related to genomics, proteomics, and systems biology. For example, some of the datasets available on Data.gov include:

- The Human Genome Project (HGP) dataset, which provides information on the human genome and its functional elements.
- The Cancer Genome Atlas (TCGA) dataset, which provides information on the genetic and epigenetic alterations in various types of cancer.
- The Gene Expression Omnibus (GEO) dataset, which provides gene expression data from a variety of organisms and experimental conditions.

To access the available datasets, visit the website (https://www.data.gov/) and use the search bar to look for datasets relevant to your research. Once you have found a dataset that you are interested in, you can download the data in a variety of formats including CSV, JSON, and XML.

It is important to note that not all datasets available may be appropriate for use in bioinformatics research. Before using any dataset, it is important to carefully review the metadata and documentation provided to ensure that the dataset is appropriate for your research question and that the data quality is sufficient.

For this module, we are using a prepared file from a previous analysis in the Warner Lab.

A Look at Heritability in *Anolis sagrei* (the Cuban Brown Anole)

Fargevieille Data Set, Warner Lab of Evolutionary Ecology, Auburn University 2019

- Set Working Directory

```
setwd("~/Desktop/Heritability")
```

- Prepare Necessary Libraries for Analysis

```
library(lme4)
```

- Create 'Masterfile' within R to Minimize the Pulled Files for Data Analysis

```
masterfile = readRDS("~/Desktop/Heritability/masterfile22.rds")
```

- Create subset of data with ONLY Juvenile Information at First Measurement, indicated in data as Juv1

```
Juv1 = masterfile %>% filter(measure == "Juv1")
```

- Creation of Models for Averaged Chromaticity (Mom and Dad) and Juv1 Chromaticity

```
chromdmean_chro.ID1 = lmer(chro~ mdmean_chro + sex + age + (1|momID),
data = Juv1, REML = FALSE)
AIC(chromdmean_chro.ID1)
summary(chromdmean_chro.ID1)

chromdmean_chro.ID2 = lmer(chro~ mdmean_chro + sex + age +(1|momID),
data = Juv1, REML = TRUE)
AIC(chromdmean_chro.ID2)
summary(chromdmean_chro.ID2)
```

Here, the response is chromaticity, with the fixed effects being the sex of the juvenile and age at time of measurement, as well as the averaged parental value for chromaticity.

Which model is best?

Interpreting Akaike Information Criterion (AIC) scores in linear mixed models (LMMs) is an important step in model selection and evaluation. AIC is a measure of the relative quality of statistical models for a given set of data. It estimates the relative amount of information lost by a given model while estimating the parameters, considering the complexity of the model. A lower AIC score indicates a better fit of the model to the data.

In R, you can obtain AIC scores for LMMs using the **AIC ( )** function. The output will provide the AIC score for each model, allowing you to compare them and choose the best one.

Let's compare the values we got from our created models:

Model: chromdmean_chro.ID1	**AIC Value: -36.92815**
Model: chromdmean_chro.ID2	**AIC Value: -10.29385**

To interpret AIC scores, it's important to keep in mind that they are not absolute measures of model quality, but rather measures of relative model performance. In other words, a lower AIC score means that a model is relatively better than another model, but it does not necessarily mean that the model is a good fit for the data.

The model with the lowest AIC score is the best model. However, a difference of two or less between AIC scores of two models indicates that both models are equally good, and you should choose the simpler one. A simpler model with fewer parameters may have a higher AIC score than a more complex model with more parameters. In such cases, you should choose the simpler model, as it is likely to have better generalization properties and avoid overfitting.

Notice that using **REML**, or restricted maximum likelihood approach in model creation altered our AIC value. Discussion question: Why might this be?

Additional Learning Material:

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. Methods in Ecology and Evolution, 5(9), 944-946.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773-795.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133-142.

We now have the *best* model; what can it tell us?

- Generate table to view model values

```
mdmean_chro_table.mdmean_chro = coef(summary(chromdmean_chro.ID2))
pval_mdmean_chro = pnorm(abs(mdmean_chro_table.mdmean_chro[, "t value"]), lower.tail = FALSE) * 2
mdmean_chro_table.mdmean_chro = cbind(mdmean_chro_table.mdmean_chro, "p value" = pval_mdmean_chro)
mdmean_chro_table.mdmean_chro
```

The resulting table includes information on the probability that a particular intercept was responsible for variation seen in the data. Look at the generated values from model 2, chromdmean_chro.ID2:

```
                Estimate  Std. Error    t value      p value
(Intercept)  1.219374615 0.478414129  2.5487847 1.080990e-02
mdmean_chro -0.043247175 0.026730714 -1.6178833 1.056878e-01
sexM         0.365874570 0.041137078  8.8940341 5.892963e-19
age         -0.001225413 0.001949498 -0.6285788 5.296249e-01
>
```

Of the intercepts, sex of the juveniles appears to be significant in the intercepts influence of chromaticity patterns observed ($p = 5.892963e^{-19}$). Age resulted in no found significance ($p = 5.296249e^{-01}$). Averaged parental chromaticity values appear to have no significant effect on the observable pattern ($p = 1.056878e^{-01}$).

Discussion Question: What aspect of *Anolis sagrei* behavior could explain a difference in dewlap chromaticity between males and females?

References

Coulson, T., et al. "Modeling adaptive and nonadaptive responses of populations to environmental change." The American Naturalist 178.6 (2011): E42-E64.

Bell, G. D., et al. "Climate change and the future of California's endemic flora." PLoS One 8.6 (2013): e66279.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Ghalambor, C. K., et al. "Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature." Nature 447.7146 (2007): 407-410.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. Springer.