HMMA307

Linear mixed effects models

Cours: Catherine Baltazar, Dalal Hanna, Jacob Ziegler *Scribes: Cassandre Lepercque*

1 Introduction

A mixed model or mixed-effects model is a statistical model containing both fixed effects and random effects.These models are useful in a wide variety of disciplines in the physical, biological and social sciences. Mixed effects models allow ecologists to overcome a number of limitations associated with traditional linear models.
In our case, we will use this model on an environmental data set. In this project, we will do a linear mixed model analysis, check its assumptions, report results, and visually represent our model.

Biological and ecological data are often messy. Usually, there is a particular structure to data and we know that relationships between variables of interest might differ depending on grouping factors like species, and more often than not sample sizes are low making it difficult to fit models that require many parameters to be estimated. Linear mixed model (lmm) are made to deal with these problems.

1.1 Objectives

1. What is a linear mixed effects model (LMM) and why should I care ?
2. How do I implement LMM's in Python ?
   (a) A priori model building and data exploration
   (b) Potential models and model selection
   (c) Model validation
   (d) Interpreting results and visualizing the model

2 What is a linear mixed effects model and why should I care ?

LMM's allow you to use all the data you have instead of using means of non-independent samples, they account for structure in your data, they allow relationships to vary by different grouping factors, and they require less parameter estimates than classical regression.

2.1 Introduction to the dataset

The dataset we will be using deals with fish trophic positions. The trophic position of an organism is the place it occupies within the food chain. Three species were selected for the study and ten individuals per species were measured (body length) in six different lakes. Here is a visual representation of the dataset.
Note : only three individuals are shown per species, but in reality there are 10 individuals per species.

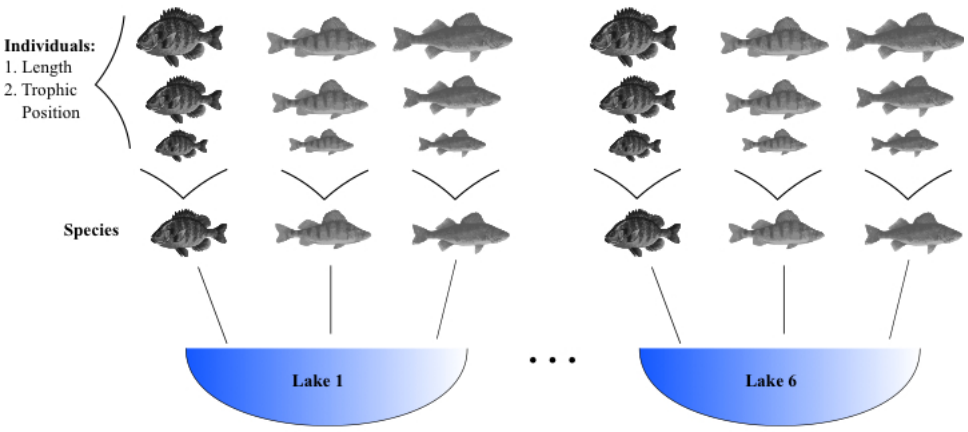

FIGURE 1 – Fish dataset.

2.2 Random and Fixed effects

There are many possible definitions, and we chose to present those we think are easier to apply.

2.2.1 Fixed effects

When a variable has a fixed effect, data is usually gathered from all it's possible levels. The person doing the analyses is also interested in making conclusions about the levels for which the data was gathered.

2.2.2 Random effects

Variables with a random effect are also called random factors, as they are only categorical variables (not continuous variables). A random effect is observed when the data only includes a random sample of the factor's many possible levels, which are all of interest. They usually are grouping factors for which you want to control the effect in your model, but are not interested in their specific effect on the response variable.

2.3 How LMM's works ?

By definition, in matrix notation, a linear mixed model can be represented as,

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

where,
- y is a known vector of observations, with mean $\mathbb{E}[y] = X\beta$,
- β is an unknown vector of fixed effects,
- u is an unknown vector of random effects, with mean $\mathbb{E}[u] = 0$,
- ϵ is an unknown vector of random errors, with mean $\mathbb{E}[\epsilon] = 0$,
- X and Z are known design matrices relating the observations y to β and u, respectively.

2.3.1 Intercepts and/or slopes are allowed to vary by lake and species

In LMM's allowing intercepts and/or slopes to vary by certain factors (random effects) simply means you assume they come from a normal distribution. A mean and standard deviation of that distribution are estimated based on your data. The most likely intercepts and slopes from that distribution are then fit by optimization (ex. maximum likelihood or restricted maximum likelihood).

2.3.2 Intercepts, slopes, and associated confidence intervals are adjusted to account for the structure of data

If a certain species or a lake is poorly represented (small sample) in the data, the model will give more importance to the grouped model to estimate the intercept and the slope of this species or this lake.
Confidence intervals of intercepts and slopes are adjusted to account for pseudoreplication based on the intraclass correlation coefficient (ICC). An adjusted sample sized based on how correlated the data within groups are.
Let's see how it goes when we have high ICC or low ICC.

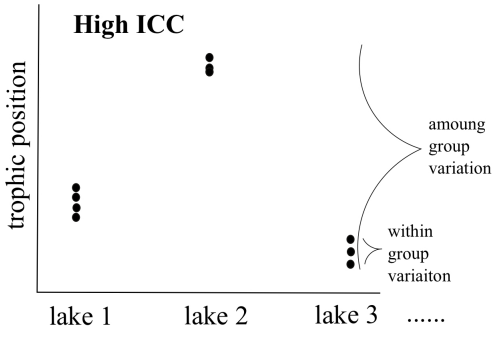

FIGURE 2 – High ICC representation

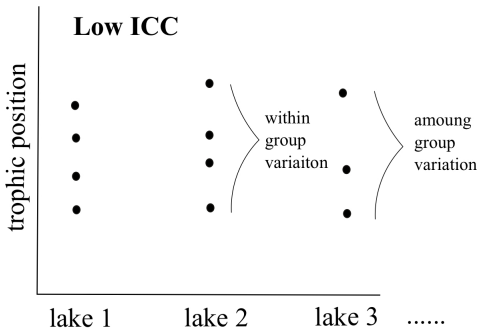

FIGURE 3 – Low ICC representation

In Figure 2, the LMM will treat points within lake more like an overall mean because they are highly correlated. Therefore, the effective sample size will be smaller leading to larger confidence intervals around your slope and intercept parameters.
In Figure 3, the LMM will treat points within lake more independently because things are less correlated

within groups compared to among groups. Therefore the effective sample size will be larger leading to smaller confidence intervals around your slope and intercept parameters.

3 How do I implement LMM's in Python ?

We now going to see how to implement LMM's in Python.

3.1 A priori model building and data exploration

We want to determine if trophic position can be predicted by body length, while taking into account variation between species and lakes. So, we have the following model,

$$TP_{ijk} \sim Length_i + Lake_j + Species_k + \epsilon,$$

where,
- TP_{ijk} is the trophic position of individual i from lake j of species k,
- ϵ are the residuals of the model.

We first look at the distribution of samples for each factor, and it's going like this :

Lake	L1	L2	L3	L4	L5	L6
	30	30	30	30	30	30
Species	S1	S2	S3			
	60	60	60			

TABLE 1 – Distribution of samples of each factor.

This data set is perfectly balanced, but mixed models can analyze unbalanced experimental designs.
Secondly, we take a look at the distribution of continuous variables. We have histograms on Figure 4 and 5. Major deviations could cause heteroskedasticity problems, to resolve that we can apply transformations to our data.
We can see in Figure 4 and 5, that our data seems great.

We want to make sur that our scale of our data is correct. If two variables in the same model have different ranges of scale, the criteria the mixed models use to come up with parameter estimates are likely to return 'convergence errors'. Z-correction standardizes your variables and adjusts for this scaling problem by placing all your variables on the same scale even if they were originally in different units :

$$z = \left(\frac{x - \mathbb{E}[x]}{\sqrt{x}} \right)$$

Because the data we are dealing with have such variable scales (length is a longer scale than trophic position), we z-correct it. We do this on trophic position and fish length, and we add it to our data set.
To know if a mixed model is necessary for our data, we need to determine whether it is important to take into account the random effect of factors that could influence the relationship that interests us (here, Lake and Species).
Let us then determine whether it is important to account for variations in "random effect" by comparing the residuals of a linear model without the random effects versus the potential random effects. Our model is : $Z_{TP} \sim Z_{Length}$.
We take the stadardized residuals from the model based on the correction Z done on fish length and trophic position.

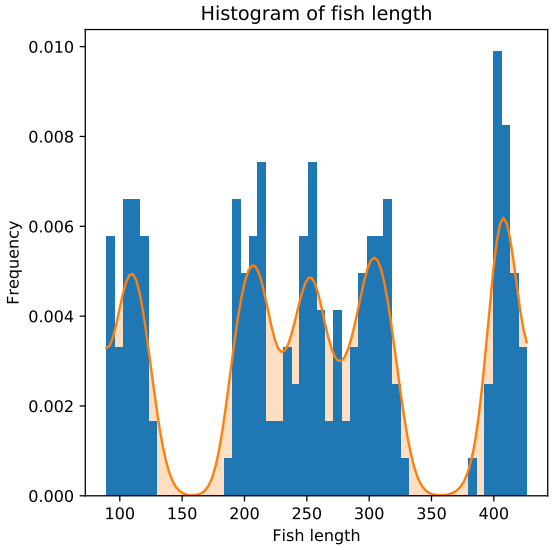

FIGURE 4 – Fish length.

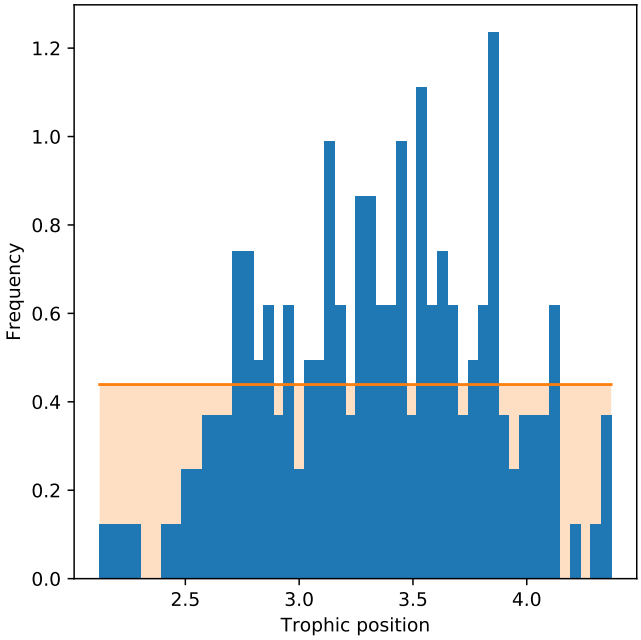

FIGURE 5 – Trophic position.

We have the boxplot representation on Figure 6 and 7.
For this model, we should keep the random effects because the standardized residuals show variation across lake and species.

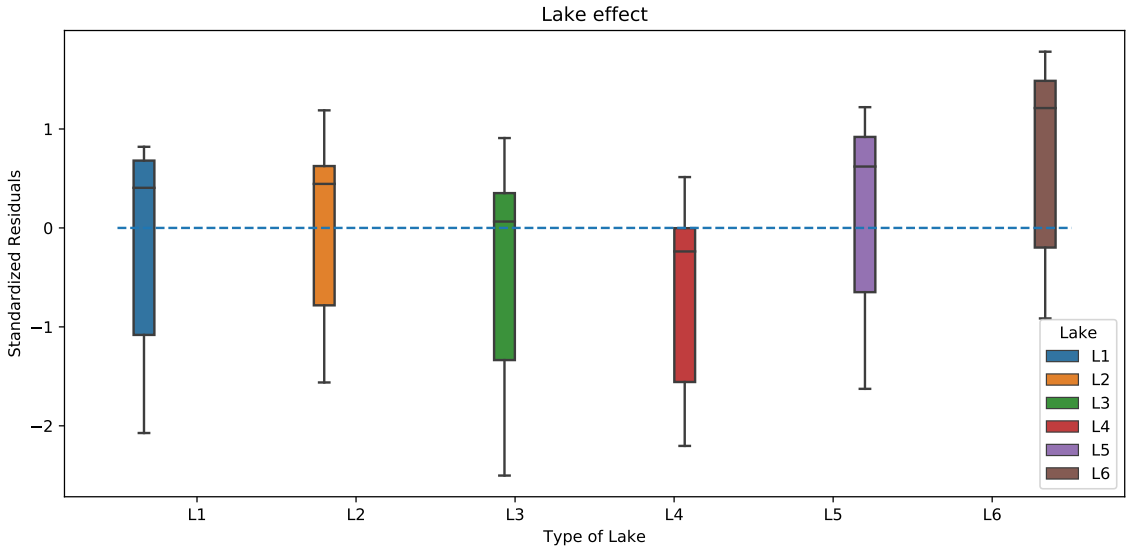

FIGURE 6 – Lake effect.

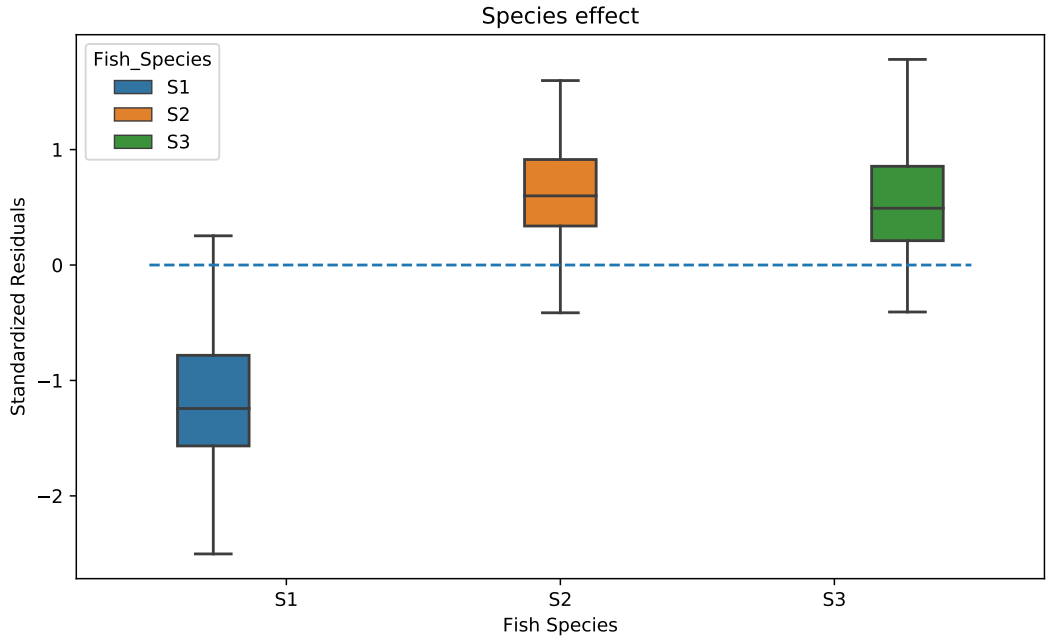

FIGURE 7 – Fish Species effect.

3.2 Potential models and model selection

We have the following a priori model :

$$TP_{ijk} \sim Length_i + Lake_j + Species_k + \epsilon$$

We are going to do a comparison of potential models to choose the better one. We want to compare them between them to select the one (those) who has (have) the greatest predictive power, and to choose correctly we are going to use models with the lowest AIC. We will, then, compare the following models.

— **Model 1 :** Linear model without random effect,
— **Model 2 :** Complete model with different intercepts,
— **Model 3 :** Complete model with differents intercepts and slopes,
— **Model 4 :** No Lake effect, random intercept only,
— **Model 5 :** No Species effect, random intercept only
— **Model 6 :** No Lake effect, random intercept and slope,
— **Model 7 :** No Species effect, random intercept and slope,
— **Model 8 :** Complete model with intercepts and slopes varying by lake,
— **Model 9 :** Complete model with intercepts and slopes varying by species.

Then, we calculate the AIC of each model and we obtain.

Models	M1	M2	M3	M4	M5	M6	M7	M8	M9
AIC	480.726	113.490	69.178	280.069	460.434	270.557	462.406	125.754	64.757

TABLE 2 – AIC of each model.

We can see that, the lowest AIC is on model 9 in the Table 2, we choose this model as the better model.

3.3 Model validation

To check the assumption of homogeneity, we must plot the predicted values against the residuals. We have the graphic bellow. The even spread of the residuals suggest that the model is a good fit for the data.

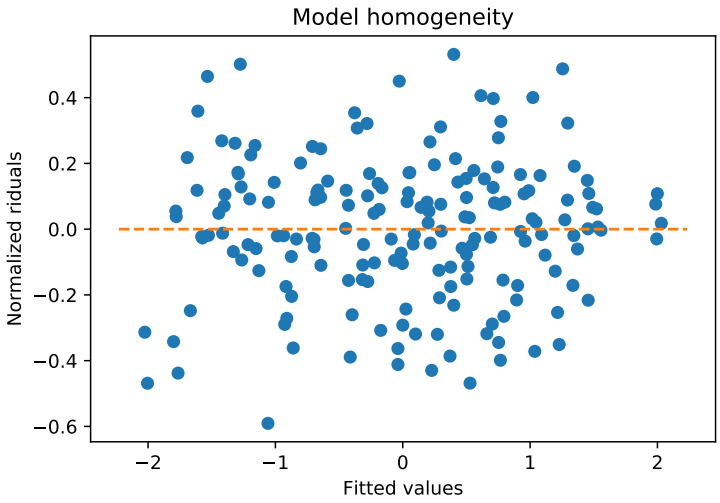

FIGURE 8 – Homogeneity.

To check the assumption of homogeneity plot residuals vs each covariate in the model.
We have the graphics on Figure 9, 11 and 12.

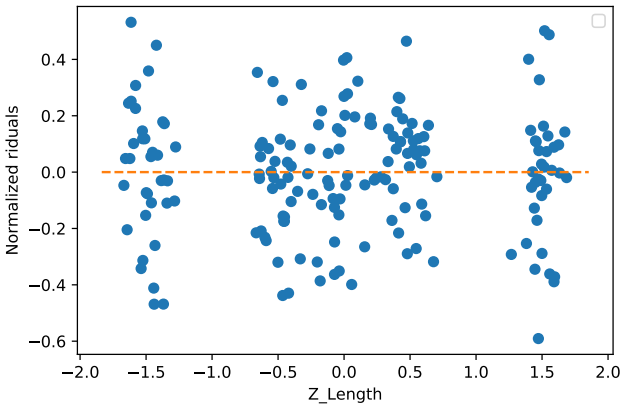

FIGURE 9 – Independence

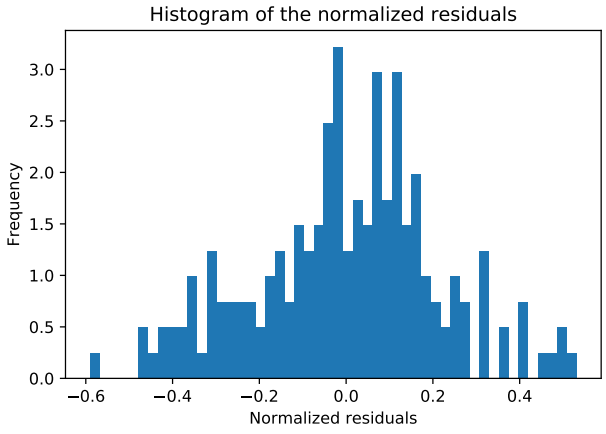

FIGURE 10 – Normality.

The equal spread above and below zero indicate that there are no homogeneity problems with these variables. It is also important to check the normality of the residues. Residues following a normal distribution indicate that the model is unbiased. The Figure 10 shows that our model is unbiased.

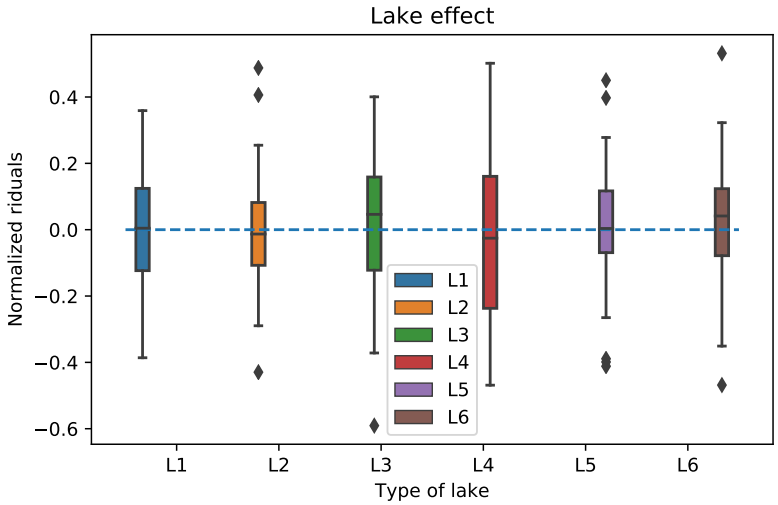

FIGURE 11 – Model validation for lake effect.

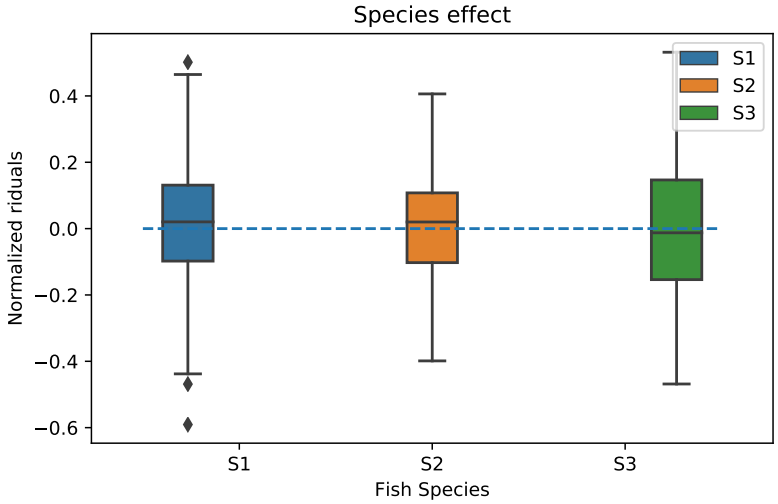

FIGURE 12 – Model validation for species effect.

3.4 Interpreting results and visualizing the model

The output is broken up into descriptions of the Random effects and Fixed effects.
To determine if the slope and, therefore, the effect of length on trophic position is significantly different from zero you first have to calculate the confidence interval (CI) for the slope parameter (estimate for Z_{Length} in the fixed effects section = 0.4223).
The CI = Standard Error of the estimate × 1.96 plus or minus the parameter estimate. If the CI overlaps with zero, then the slope is not significantly different from zero at the 0.05 level. In our case, the Intercept is -0.09. So, we have,

— upper CI = 0.4223 + 0.09*1.96 = 0.5987,
— lower CI = 0.4223 - 0.09*1.96 = 0.2459.

As *zero* isn't in the interval, we can say that the Z_{Length} slope is significantly different from 0.
To visualize this model you must obtain the coefficients (intercept and slope) of each component of the model. Overall our group level model coefficients can be found in the summary of the model in the fixed effects section. We have the following graphic.

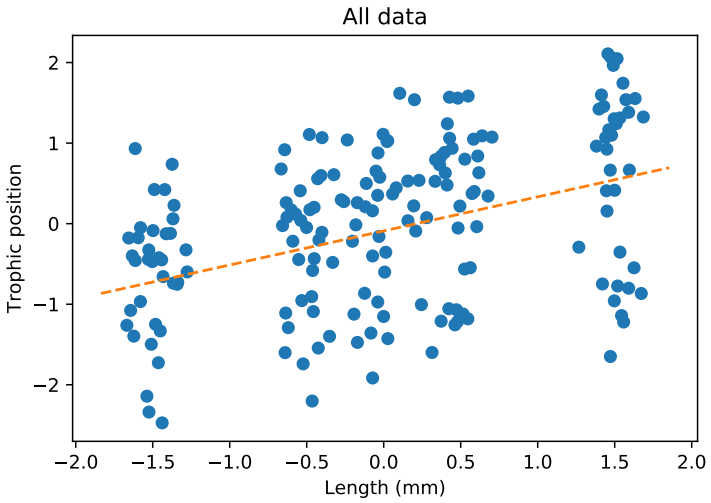

FIGURE 13 – Model visualization of all the dataset.

4 Conclusion

In our case, the trophic position increases with the size of the fish.

Mixed models are really good at accounting for variation in ecological data while not loosing too many degrees of freedom.
We have covered only a small portion of what LMM's can do. Maybe with other dataset, we will use different thing to have a great model and interpretation.

Références

[CBZ19] Dalal Hanna Catherine Baltazar and Jacob Ziegler. Workshop 6 : Linear mixed effects models. https://wiki.qcbs.ca/r_atelier6#qcbs_r_workshops, 2019.

[Lep20] Cassandre Lepercque. Linear mixed effects model, python code. https://github.com/cassandrelepercque/HMMA_307-Project/blob/main/code.py, 2020.

[Sal20] Joseph Salmon. Advanced linear models hmma307. http://josephsalmon.eu/HMMA307.html, 2020.

[Wik] Wikipedia. https://en.wikipedia.org/wiki/Mixed_model.