



# Atelier 6: modèles linéaires à effets mixtes

CSBQ R Série d'Atelier

Centre des Sciences de la Biodiversité du Québec



# À propos de cet atelier



# Packages requis

- `ggplot2`
- `lme4`
- `AICcmodavg`

```
install.packages(c('ggplot2', 'lme4', 'AICcmodavg'))
```

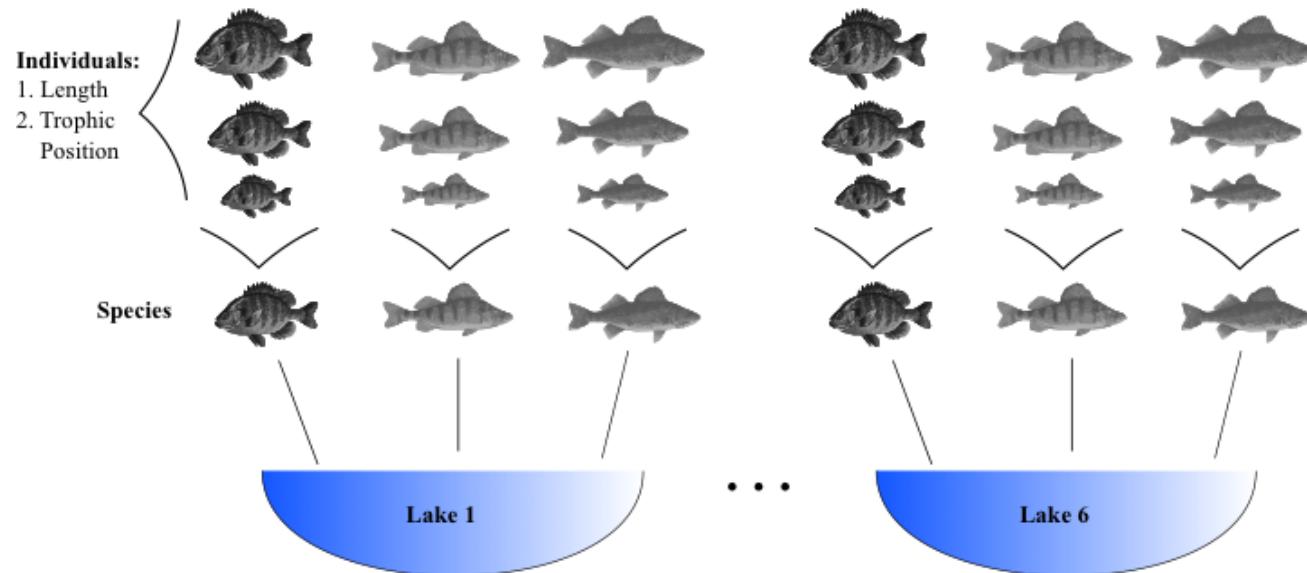
# Objectifs d'apprentissage

1. Comprendre ce que sont les modèles linéaires à effets mixtes (MLMs)
2. Comprendre leur pertinence notamment pour les sciences de la biodiversité
3. Apprendre à utiliser les MLMs dans R
  - Exploration des données
  - Construction du modèle *\*a priori\**
  - Coder les modèles potentiels
  - Sélection du meilleur modèle
  - Validation du modèle
  - Interprétation et visualisation des résultats

# Question

## Est-ce que la position trophique des poissons augmente avec leur taille?

Pour répondre, 3 espèces ont été sélectionnées et dix individus par espèce ont été mesurés (longueur corporelle) dans six lacs différents.

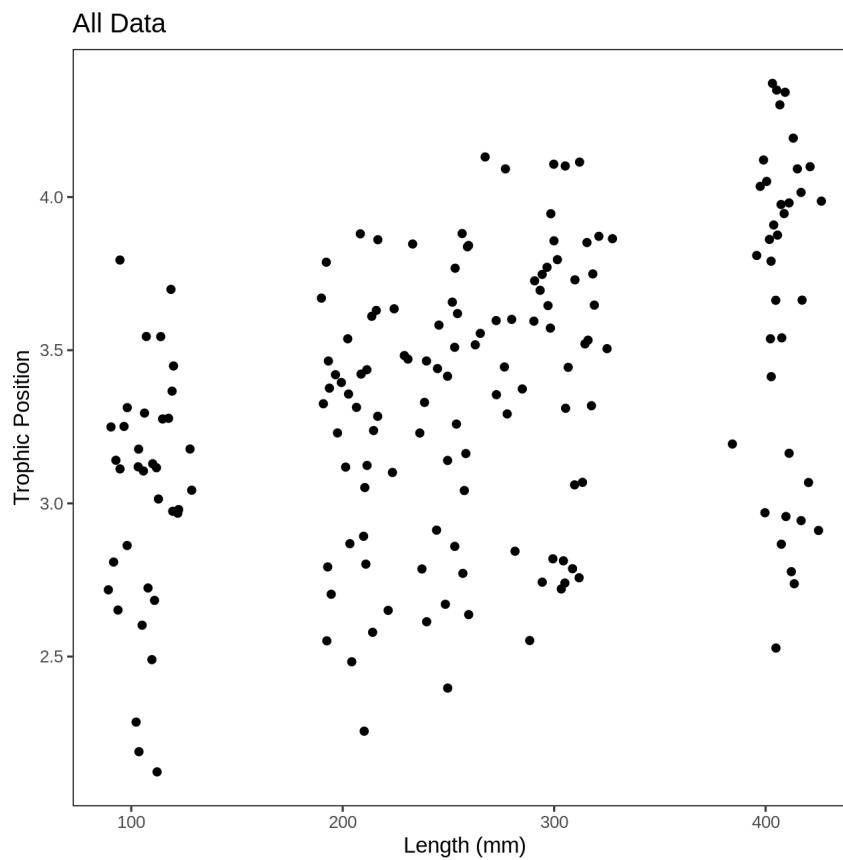


# Familiarisez vous avec le jeu de données

1. Ouvrez le jeu de données,
2. Ouvrez le script de l'atelier dans R,
3. Visualisez la relation entre taille et position trophique.

```
data ← read.csv('data/qcbs_w6_data.csv')
head(data)
#   Lake Fish_Species Fish_Length Trophic_Pos
# 1 L1      S1       105.1501    2.602388
# 2 L1      S1       194.5708    2.703522
# 3 L1      S1       294.3636    2.742878
# 4 L1      S1       413.5295    2.737743
# 5 L1      S1       237.4739    2.785936
# 6 L1      S1       107.9315    2.723862
```

# Visualisation avec toutes les données



# Visualisation de l'ensemble des données

```
# thème simplifié
fig ← theme_bw() + theme(panel.grid.minor=element_blank(),
  panel.grid.major=element_blank(), panel.background=element_blank()) +
  theme(strip.background=element_blank(), strip.text.y = element_text())
  theme(legend.background=element_blank()) +
  theme(legend.key=element_blank()) +
  theme(panel.border = element_rect(colour="black", fill=NA))

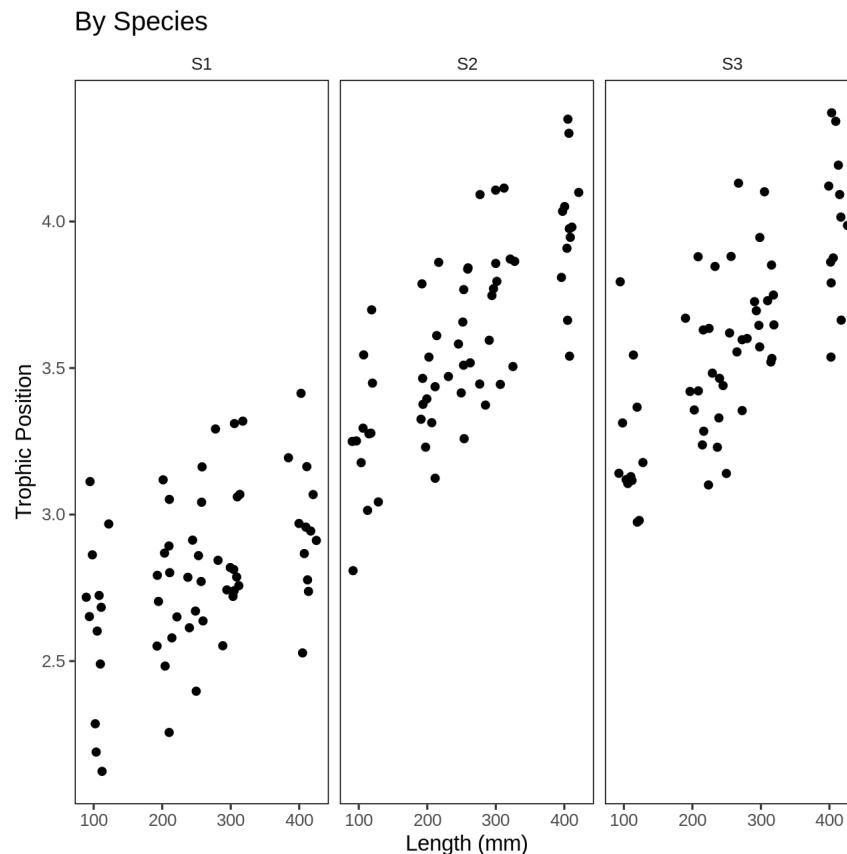
# Faites les trois graphiques suivants pour explorer les données
plot ← ggplot(aes(Fish_Length,Trophic_Pos),data=data)

# Graphique 1 - Toutes les données
plot + geom_point() + xlab("Length (mm)") + ylab("Trophic Position") + la
```

# Visualisation par espèce

# Graphique 2 - Par espèce

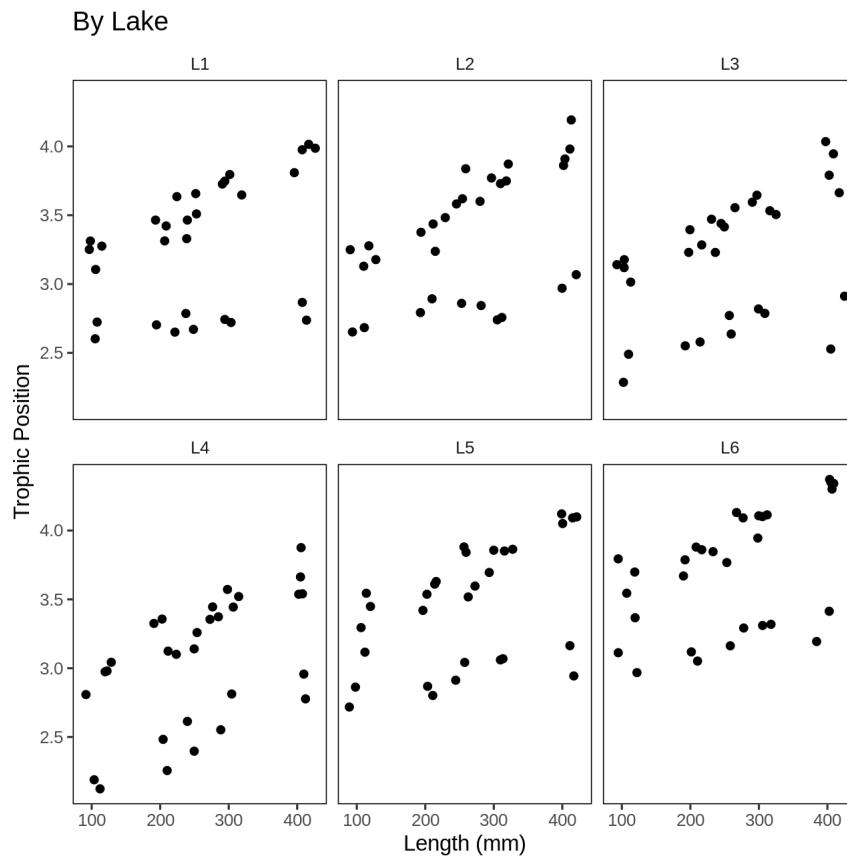
```
plot + geom_point() + facet_wrap(~ Fish_Species) + xlab("Length (mm)") +  
  labs(title="By Species") + fig
```



# Visualisation par Lake

# Graphique 3 – Par lac

```
plot + geom_point() + facet_wrap(~ Lake) + xlab("Length (mm)") + ylab("Trophic Position")  
  labs(title="By Lake") + fig
```



# Pourquoi choisir un MLM?

## Discussion de groupe

- Est-ce qu'on s'attend à ce que, pour toutes les espèces, la position trophique augmente avec la longueur corporelle?
  - Exactement de la même façon?
- Est-ce qu'on s'attend à ce que ces relations soient pareilles entre les lacs?
  - Comment pourraient-elles différer?

# Pourquoi choisir un MLM?

**Comment pourrions-nous analyser ces données?**

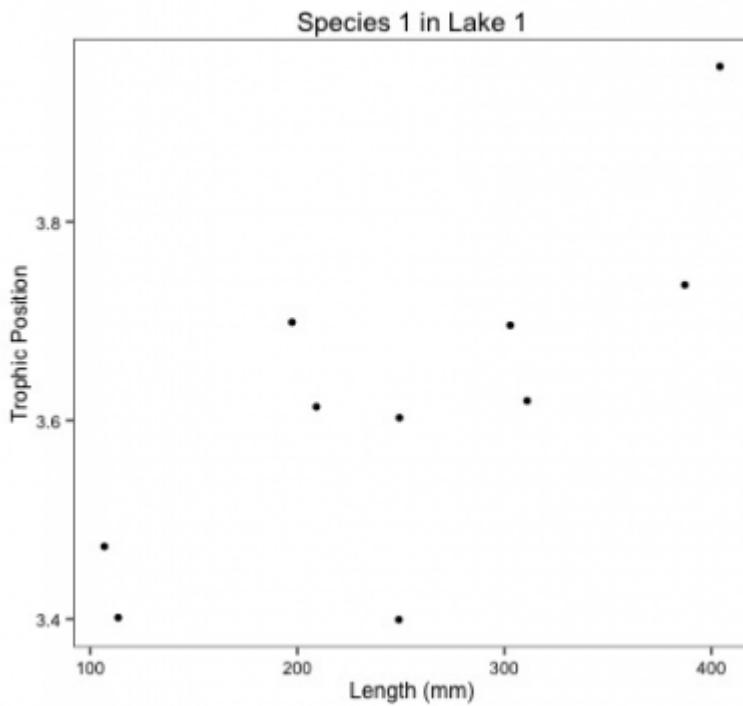
**Option 1. Séparer:**

- Faire une analyse séparée pour chaque espèces et chaque lac

**Option 2. Regrouper:**

- Faire une seule analyse en ignorant les variables espèces et lac

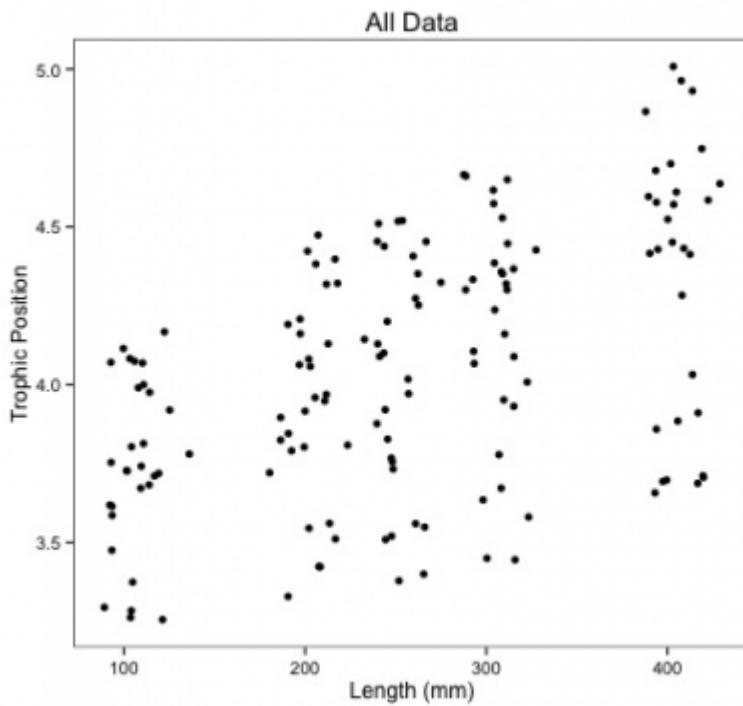
# Pourquoi choisir un MLM?



## Option 1. Séparer

- Estime 6 intercepts et 6 pentes pour chaque espèces (i.e. 6 lacs)
- Taille d'échantillon  $n = 10$  pour chaque analyse (i.e. 10 poissons/espèces/lac)
- Peu de chances de détecter un effet a cause de la faible taille d'échantillon  $n$

# Pourquoi choisir un MLM?



## Option 2. Regrouper:

- Très grande taille d'échantillon!
- Et la pseudoreplication? (les poissons d'un même lac et d'une même espèce sont corrélés).
- Beaucoup de bruit! Une partie doit être due aux effets de l'espèce et du lac.

# Pourquoi choisir un MLM?

- Pour **notre question**, on veut seulement savoir s'il y a un **effet général de la longueur corporelle sur la position trophique**,
- Ceci pourrait varier faiblement par espèce à cause de différents taux de croissance et/ou par lac à cause de différences dans la disponibilité de nourriture. **On ne s'intéresse pas directement à ces facteurs non mesurés, mais on doit contrôler leur effet dans le modèle.**

Les MLMs sont un **compromis entre séparer et regrouper**. Ils:

1. Prendre entre compte de la variabilité spécifique pour chaque espèce et chaque lac (séparer) mais en calculant moins de paramètres;
2. Utilisent toutes les données disponibles (regrouper) tout en contrôlant les différences entre les lacs et les espèces (pseudo-replication).

# Pourquoi choisir un MLM?

## Effet fixe VS effet aléatoire

Dans la littérature des MLMs, vous rencontrerez souvent ces termes souvent. Il existe plusieurs façons de les présenter et nous vous présenterons ici celles que nous trouvons les plus faciles à appliquer.

### Effet fixe

- Les données proviennent de tous les niveaux possibles d'un facteur (variable qualitative),
- On souhaite émettre des conclusions à propos des niveaux du facteur d'où les données proviennent.

### Effet aléatoire

- Seulement des variables qualitatives = facteur aléatoire;
- Permet de structurer le processus d'erreur.

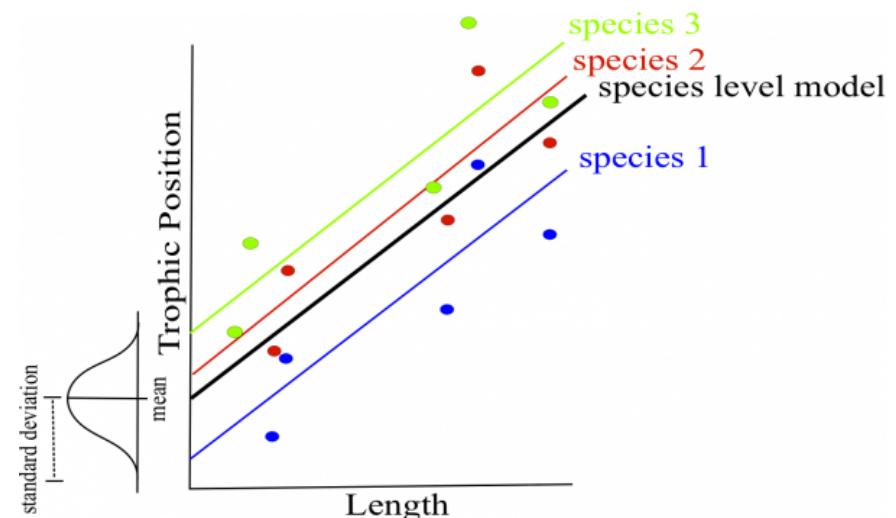
# Pourquoi choisir un MLM?

## Comment fonctionnent les MLMs?

- A.** Permet aux intercepts et/ou aux pentes d'être considérés comme propre à une population donnée (**effet aléatoire**), e.g. par lac et/ou par espèce
- B.** Les intercepts, les pentes et leur intervalle de confiance sont ajustés pour **prendre en compte la structure des données.**

# Effet aléatoire sur l'intercept

On fait la supposition que les intercepts proviennent d'une distribution normale et on a seulement besoin d'estimer la moyenne (intercept général) et l'écart type de la distribution normale (effet aléatoire) au lieu d'estimer un intercept par espèce (ce qui rajoute 2 paramètres).



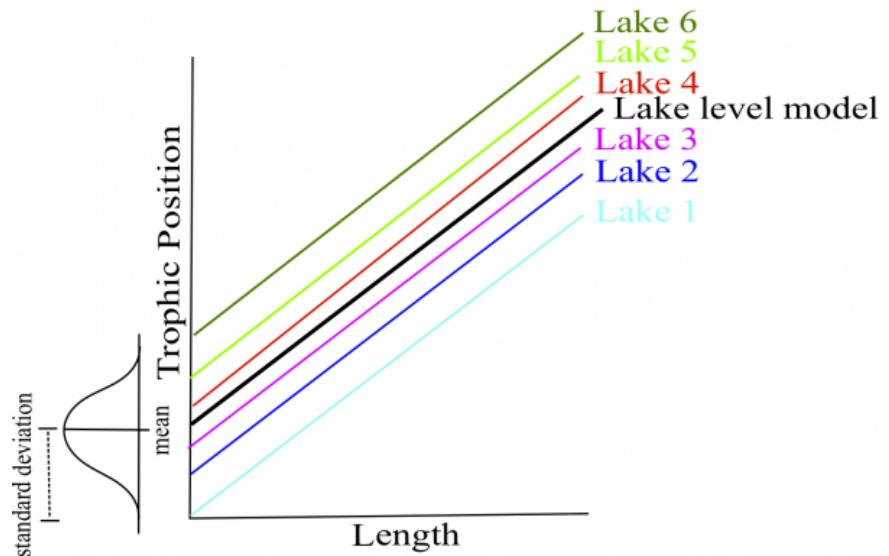
Au lieu d'estimer un intercept par espèce (trois paramètres pour 3 espèces) on estime un intercept générale et un effet aléatoire soit 2 paramètres. Avec  $n$  espèce, dans le premier cas on estime  $n - 1$  paramètres alors qu'on reste à 2 paramètres dans le second cas!

# Effet aléatoire sur l'intercept

Même principe pour les lacs

Estime 2 paramètres (moyenne et écart-type) au lieu de 6 intercepts.

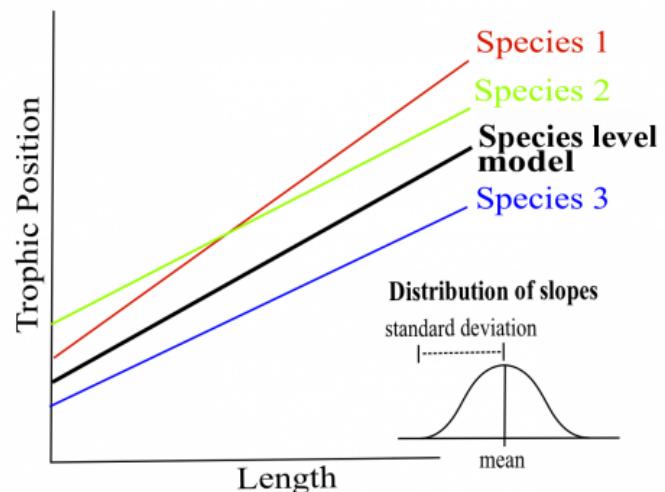
Cela économise des degrés de liberté (moins d'estimation de paramètres sont nécessaires)



# Effet aléatoire sur la pente

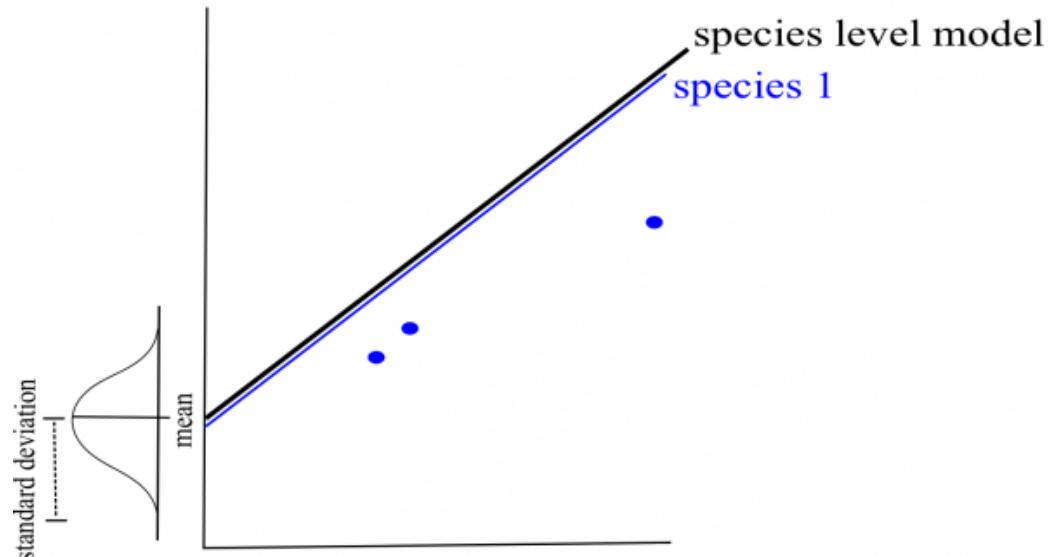
Le même principe s'applique aux pentes qui varient selon un facteur donné, juste plus difficile à visualiser.

Comme pour les intercepts, seuls la moyenne et l'écart-type des pentes sont estimés au lieu de trois pentes distinctes.



# Tenir compte de la structure des données

Si une certaine espèce ou un lac est peu représenté (faible  $n$ ) dans les données, le modèle va accorder plus d'importance au modèle groupé pour estimer l'intercept et la pente de cette espèce ou de ce lac.

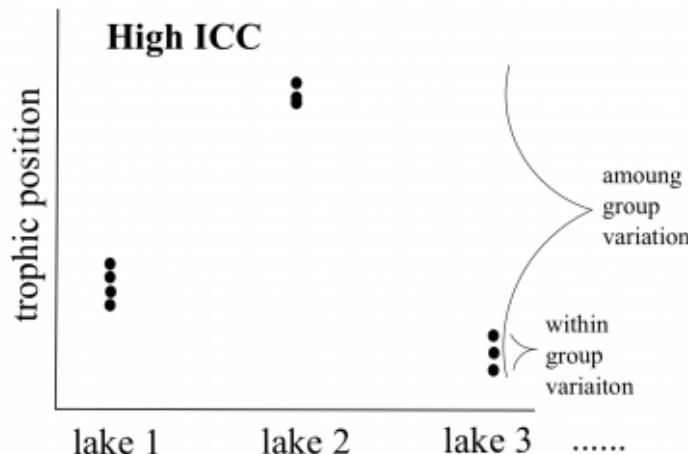


# Tenir compte de la structure des données

- Les intervalles de confiance des intercepts et pentes sont ajustés pour tenir compte de la pseudo-replication basée sur le **coefficent de corrélation intra-classe (CCI)**.
- Combien de variation y a-t-il dans chaque groupe VS entre les groupes ?

# Tenir compte de la structure des données

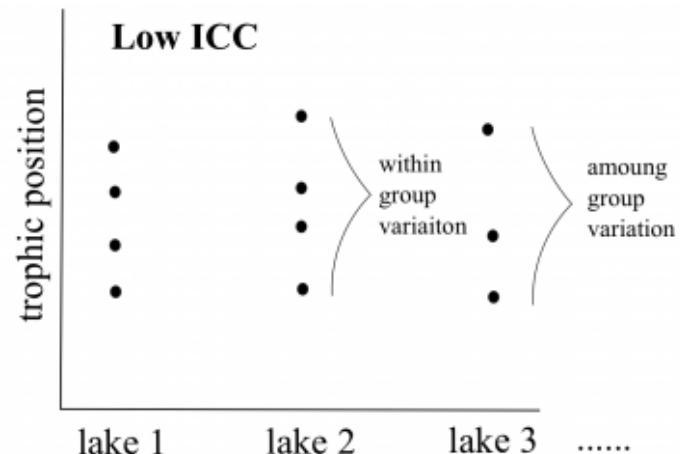
## CIC élevé



les points provenant d'un même lac sont traités comme une seule observation car très corrélés

➔ petite taille effective de l'échantillon et grands intervalles de confiance pour la pente et l'intercept.

## CIC faible



les points provenant d'un même lac sont traités indépendamment car peu corrélés

➔ grande taille effective de l'échantillon et petits intervalles de confiance pour la pente et l'intercept.

# Question / défi



**Comment le CIC et l'intervalle de confiance seront affectés dans ces deux scénarios ?**

**Q1.** Les positions trophiques des poissons ne varient pas entre les lacs

**Q2.** Les positions trophiques des poissons sont similaires dans les lacs mais différentes entre les lacs



# Solution

**Q1.** La position trophique ne varie pas entre les lacs?

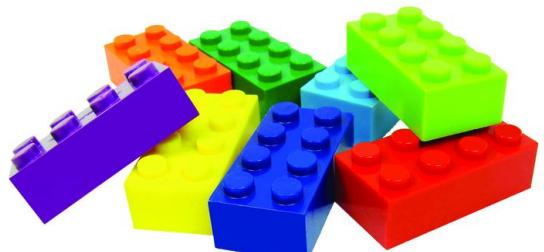
**R1. CIC faible, petits intervalles de confiances**

**Q2.** La position trophique est similaire dans un lac mais différente entre les lacs ?

**R2. CIC élevé, grands intervalles de confiance**

# Comment implémenter un MLM dans R ?

Étape 1: Construction du modèle *a priori* et exploration des données



**Étape 2:** Coder les modèles potentiels et sélection du meilleur modèle

**Étape 3:** Validation du modèle

**Étape 4:** Interprétation et visualisation des résultats

# Étape 1 - exploration des données

- Modèle basé sur connaissance *a priori*:
  - Nous voulons déterminer si la position trophique peut être prédite par la longueur corporelle, tout en prenant en compte la variation entre les espèces et les lacs
  - Donc nous voulons un modèle qui ressemble a ceci:

$$PT_{ijk} \sim Longueur_i + Lac_j + Espèce_k + \varepsilon$$

# Étape 1 - exploration des données

## Les données ont-elles la bonne structure?

```
data ← read.csv('data/qcbs_w6_data.csv')
str(data)
# 'data.frame': 180 obs. of 4 variables:
# $ Lake       : Factor w/ 6 levels "L1", "L2", "L3", ..: 1 1 1 1 1 1 1 1
# $ Fish_Species: Factor w/ 3 levels "S1", "S2", "S3": 1 1 1 1 1 1 1 1 1 1
# $ Fish_Length : num 105 195 294 414 237 ...
# $ Trophic_Pos : num 2.6 2.7 2.74 2.74 2.79 ...
```

**Il est recommandé de faire le ménage de votre espace de travail (`rm.list()`) avant de construire un modèle.**

# Étape 1 - exploration des données

Regardez la distribution des échantillons pour chaque facteur:

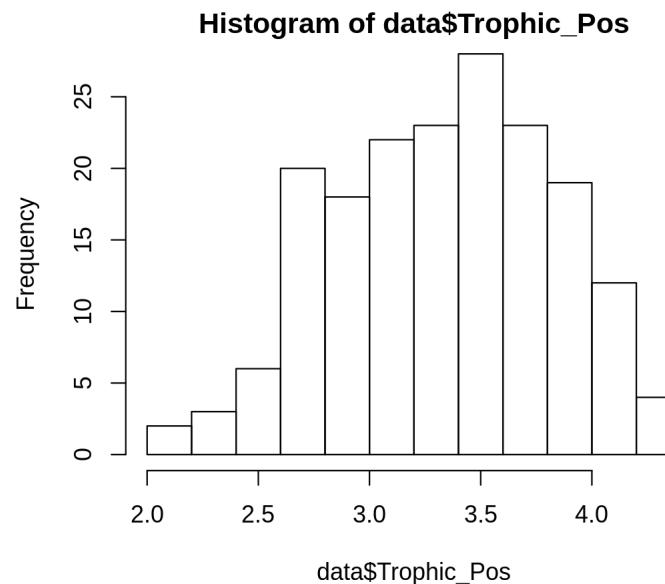
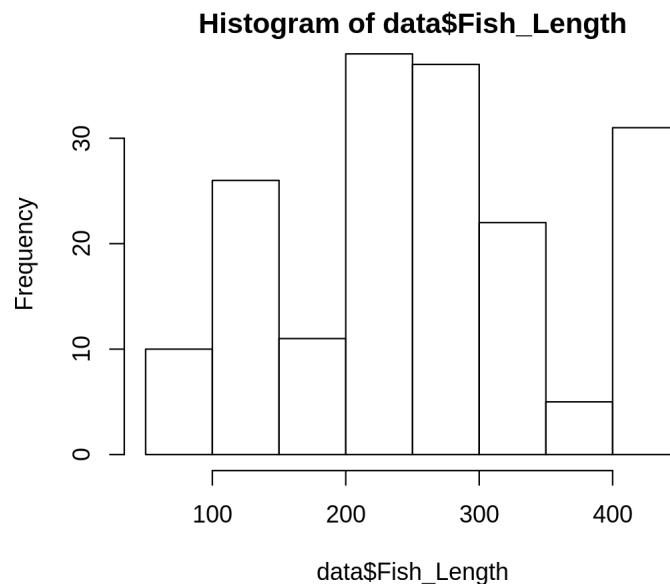
```
table(data$Lake)
#
# L1 L2 L3 L4 L5 L6
# 30 30 30 30 30 30
table(data$Fish_Species)
#
# S1 S2 S3
# 60 60 60
```

Ce jeu de données est parfaitement équilibré, mais les **modèles mixtes peuvent analyser les plans expérimentaux non équilibrés**, comme c'est souvent le cas en écologie!

# Étape 1 - exploration des données

Regardez la distribution des variables continues

```
par(mfrow=c(1,2), mar = c(4,4,1,1))
hist(data$Fish_Length)
hist(data$Trophic_Pos)
```



Des déviations majeures pourraient causer des problèmes d'hétérosécédasticité. Si nécessaire, faites des transformations. Dans ce cas-ci, **les données semblent correctes**.

# Étape 1 - exploration des données

## Vérification de la colinéarité entre vos variables explicatives

Le problème avec les prédicteurs colinéaires est simplement qu'ils expliquent la même chose, alors leur effet sur la variable réponse sera confondu dans le modèle.

Dans cet exemple, il n'y a pas de risque de colinéarité avec seulement une variable continue. Si vous aviez une autre variable continue (Var2), une façon simple de vérifier la colinéarité est:

```
plot(data)
```

```
cor(var1, var2)
```



# Question / défi

Quelles mesures supplémentaires aurions-nous pu prendre sur le terrain et qui auraient pu être fortement corrélées avec la longueur corporelle?

Un exemple est la masse du poisson – c'est une variable fortement corrélée avec la longueur du poisson. Par conséquent, nous ne voulons pas inclure ces deux variables dans le même modèle.

# Étape 1 - exploration des données

## Considérez l'échelle de vos données

- Si deux variables dans un même modèle ont des échelles très différentes, il est probable que le modèle mixte indique un problème de convergence en essayant de calculer les paramètres.
- La **correction Z** standardise les variables et résout ce problème (fonction `scale()` dans R) :

$$z = \frac{(x - \bar{x})}{\sqrt{\text{var}(x)}}$$

# Étape 1 - exploration des données

## Considérez l'échelle de vos données

- Longueur corporelle → Longue échelle
- Position trophique → Courte échelle

# Étape 1 - exploration des données

## Considérez l'échelle de vos données

- Parce que nos données ont des échelles très différentes, on applique la **correction Z**

*#Longueur corrigée, "à la main"*

```
data$Z_Length ← (data$Fish_Length-mean(data$Fish_Length))/sd(data$Fish_L
```

*#Position trophique corrigée, avec scale*

```
data$Z_TP← scale(data$Trophic_Pos)
```

# Étape 1 - exploration des données

Pour savoir si un modèle mixte est nécessaire pour vos données, vous devez déterminer s'il est important de prendre en compte l'effet aléatoire de facteurs qui pourraient influencer la relation qui vous intéresse (dans notre cas, lac et espèce)

Nous pouvons le faire en :

1. Cr  ant un mod  le lin  aire sans les facteurs qui pourraient avoir un effet al  atoire
2. Calculant les r  sidus de ce mod  le lin  aire
3. Produisant un graphique de la valeur des r  sidus en fonction des niveaux des facteurs potentiellement al  atoires

# Étape 1 - exploration des données

1. Créer un modèle linéaire sans les facteurs

```
lm.test <- lm(Z_TP ~ Z_Length, data = data)
```

1. Calculer les résidus de ce modèle linéaire

```
lm.test.resid <- rstandard(lm.test)
```

# Exploration des données

Représentez graphiquement la valeur des résidus en fonction des niveaux des facteurs

```
par(mfrow=c(1,2))

plot(lm.test.resid ~ data$Fish_Species,
     xlab = "Species", ylab = "Standardized residuals")

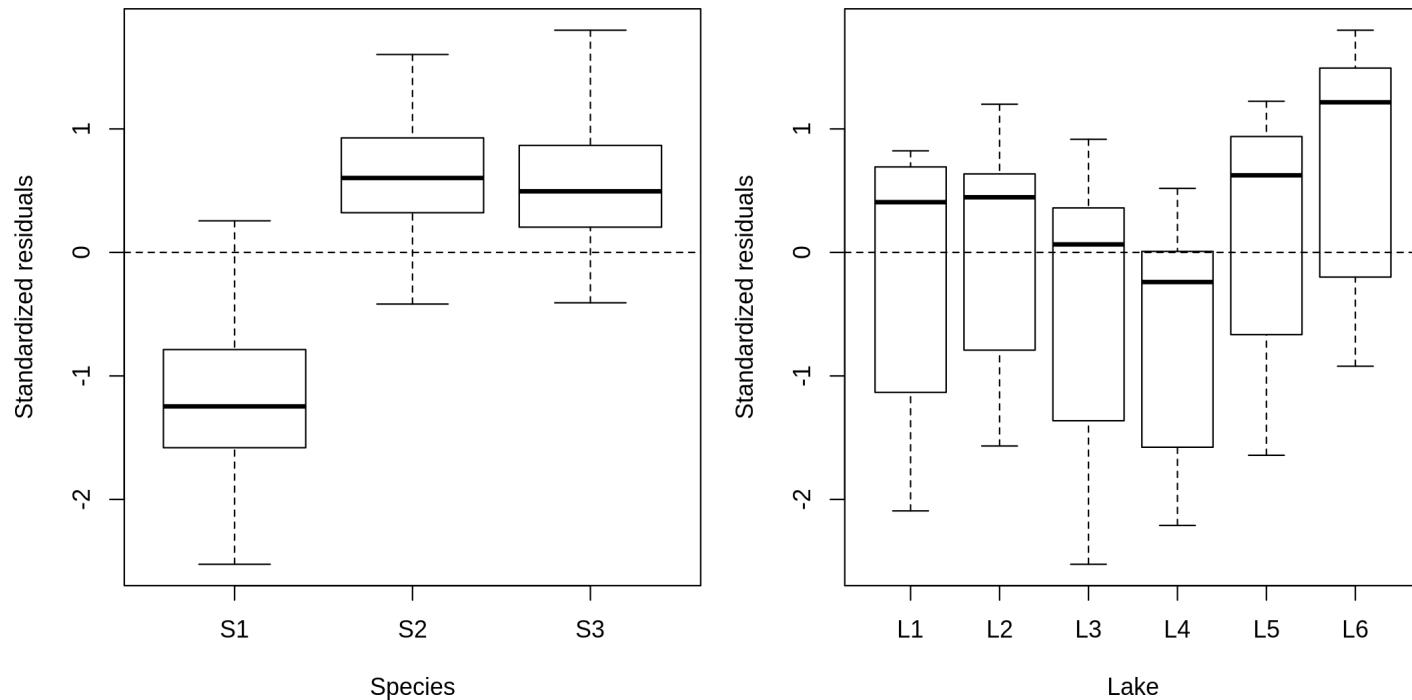
abline(0, 0, lty = 2)

plot(lm.test.resid ~ data$Lake,
     xlab = "Lake", ylab = "Standardized residuals")

abline(0, 0, lty = 2)
```

# Exploration des données

Représentez graphiquement la valeur des résidus en fonction des niveaux des facteurs



Ces patrons suggèrent qu'il y a de la variance résiduelle qui pourrait être expliquée par ces facteurs, et ils devraient donc être inclus dans le modèle

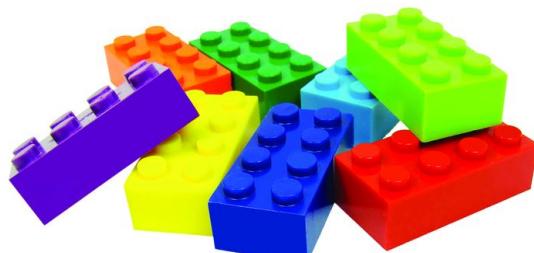
# Comment implémenter un MLM dans R ?

Étape 1: Construction du modèle *a priori* et exploration des données

Étape 2: Coder les modèles potentiels et sélection du meilleur modèle

Étape 3: Validation du modèle

Étape 4: Interprétation et visualisation des résultats



# Étape 2 - coder les modèles

Traduisons notre modèle...

$$PT_{ijk} \sim Longueur_i + Lac_j + Espèce_k + \varepsilon$$

... En code R

```
library(lme4)
lmer(Z_TP ~ Z_Length + (1 | Lake) + (1 | Fish_Species),
      data = data, REML = TRUE)
```

- `lmer` → fonction "linear mixed model" du package `lme4`
- `(1 | Lake)` → indique que les intercepts peuvent varier
- `REML = TRUE` → méthode d'estimation

# Note à propos de la méthode d'estimation

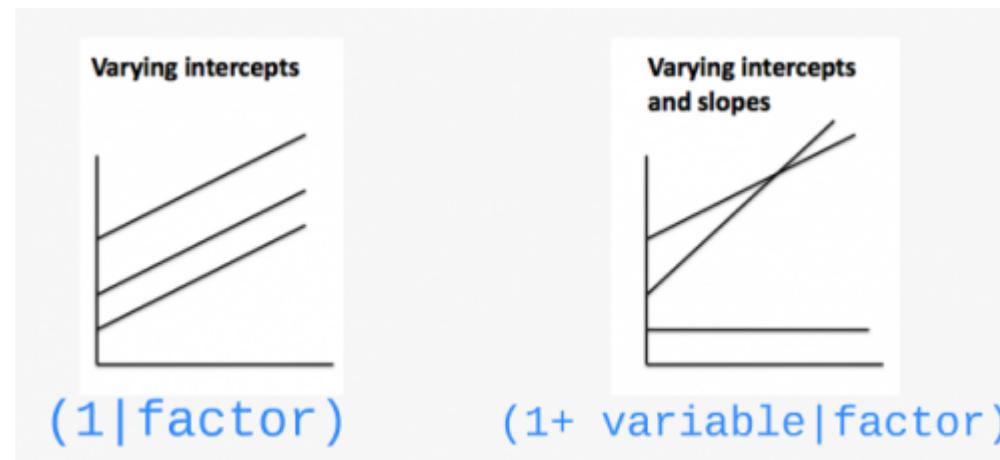
REML (Restricted Maximum Likelihood) est la méthode par défaut dans la fonction `lmer` (voir `?lmer`).

Il est à noter que l'estimateur de l'écart-type du maximum de vraisemblance (ML, pour Maximum Likelihood) est biaisé d'un facteur  $(n - 2)/n$ . La méthode REML corrige ce biais.

- On devrait comparer les **modèles d'effets aléatoires nichés avec REML**
- Tandis qu'on devrait comparer les **modèles nichés à effets fixes avec ML**

# Étape 2 - coder et sélectionner le meilleur modèle

Comment faire si on souhaite que la pente puisse varier ?



# Étape 2 - coder et sélectionner le meilleur modèle

## Plus généralement

- $(1 \mid \text{Lake})$  effet aléatoire par lac sur l'ordonnée à l'origine
- $(1 + \text{Z_Length} \mid \text{Lake})$  effet aléatoire par lac sur la pente à l'origine et la pente (NB:  $(\text{Z_Length} \mid \text{Lake})$  donne le même résultat)
- $(-1 + \text{Z_Length} \mid \text{Lake})$  pour avoir uniquement l'effet aléatoire sur la pente
- $(1 \mid \text{Lake}) + (1 \mid \text{Species})$  pour des effets aléatoires croisés
- $(1 \mid \text{Lake}:\text{Fish_Species})$  pour utiliser l'interaction entre 2 facteurs groupant
- si votre jeu de données inclus des effets nichés, vous pouvez utiliser  $/$  pour les déclarer, e.g.  $(1 \mid \text{facteur1} / \text{facteur2})$  si  $\text{facteur2}$  est niché dans  $\text{facteur1}$  (voir 

# Défi 3



Réécrivez le code suivant de façon à ce que les **pentes** de la relation position trophique en fonction de longueur corporelle **variant par lac et par espèces**:

```
lmer(Z_TP ~ Z_Length + (1 | Lake) + (1 | Fish_Species),  
      data = data, REML = TRUE)  
# Linear mixed model fit by REML ['lmerMod']  
# Formula: Z_TP ~ Z_Length + (1 | Lake) + (1 | Fish_Species)  
# Data: data  
# REML criterion at convergence: 72.4662  
# Random effects:  
# Groups           Name        Std.Dev.  
# Lake             (Intercept) 0.4516  
# Fish_Species     (Intercept) 0.9301  
# Residual          0.2605  
# Number of obs: 180, groups: Lake, 6; Fish_Species, 3  
# Fixed Effects:  
# (Intercept)    Z_Length  
# 9.752e-14     4.198e-01
```

# Solution



```
lmer(Z_TP ~ Z_Length + (1 + Z_Length | Lake) + (1 + Z_Length | Fish_Speci
    data = data, REML = TRUE)
# Linear mixed model fit by REML ['lmerMod']
# Formula:
# Z_TP ~ Z_Length + (1 + Z_Length | Lake) + (1 + Z_Length | Fish_Species)
# Data: data
# REML criterion at convergence: 20.579
# Random effects:
# Groups           Name        Std.Dev. Corr
# Lake             (Intercept) 0.45336
#                  Z_Length    0.02373 -0.82
# Fish_Species     (Intercept) 0.92359
#                  Z_Length    0.15609  1.00
# Residual
# Number of obs: 180, groups: Lake, 6; Fish_Species, 3
# Fixed Effects:
# (Intercept)    Z_Length
# -0.0009031    0.4223750
# convergence code 0; 1 optimizer warnings; 0 lme4 warnings
```

## Étape 2 - coder et sélectionner le meilleur modèle

- Pour déterminer si vous avez construit le meilleur modèle mixte base sur vos connaissances a priori, vous devez comparer ce modèle a priori aux autres modèles alternatifs,
- Avec le jeu de données sur lequel vous travaillez, il y a plusieurs modèles alternatifs qui pourraient mieux correspondre à vos données.



## Défi 4

Faites une liste de 7 modèles alternatifs qui pourraient être comparés à celui-ci:

```
lmer(Z_TP ~ Z_Length + (1 | Lake) + (1 | Fish_Species),  
      data = data, REML = TRUE)
```

Note: Si nous avions différents effets fixes entre les modèles, nous aurions dû indiquer `REML=FALSE` pour les comparer avec un méthode de vraisemblance comme l'AIC. Ici, vous devez rapporter les estimations des paramètres du "meilleur" modèle en utilisant `REML=TRUE`



# Solution

- Nous allons aussi construire le **modèle linéaire de base** `lm()` parce qu'il est toujours utile de voir la variation dans les valeurs de AICc.

```
M0 ← lm(Z_TP ~ Z_Length, data = data)
```

- Par contre, pour comparer ce modèle aux MLMs, il est important de **changer la méthode d'estimation à ML (REML=FALSE)** parce que `lm()` n'utilise pas la même méthode d'estimation que `lmer()`
  - Démontrer que les résultats de la méthode des moindres carrés (least squares) est équivalente aux résultats de la méthode ML pour les modèles linéaires de bases!

# Solution

```
# Modèle linéaire de base
M0 ← lm(Z_TP ~ Z_Length, data = data)
# modèle complet avec variation des intercepts
M1 ← lmer(Z_TP ~ Z_Length + (1 | Fish_Species) + (1 | Lake), data = data)
# modèle complet avec variation des intercepts et de pentes
M2 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 + Z_Length | Lake),
           data = data, REML = FALSE)
# Pas d'effet lac, les intercepts varient par espèces
M3 ← lmer(Z_TP ~ Z_Length + (1 | Fish_Species), data = data, REML = FALSE)
# Pas d'effet espèces, les intercepts varient par lac
M4 ← lmer(Z_TP ~ Z_Length + (1 | Lake), data = data, REML = FALSE)
# Pas d'effet de lac, les intercepts et les pentes varient par espèces
M5 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species), data = data,
           REML = FALSE)
# Pas d'effet de l'espèces, les intercepts et les pentes varient par lac
M6 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Lake), data = data, REML = FALSE)
# modèle complet, variation d'intercept et pente par lac
M7 ← lmer(Z_TP ~ Z_Length + (1 | Fish_Species) + (1 + Z_Length | Lake),
           data = data, REML = FALSE)
# modèle complet, variation d'intercept et pente par espèces
M8 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 | Lake),
```

# Coder et sélectionner le meilleur modèle

- Maintenant que nous avons une liste de modèles potentiels, nous voulons les comparer entre eux pour sélectionner celui(ceux) qui a(ont) le plus de pouvoir prédictif
- Les modèles peuvent être comparés en utilisant la fonction `AICc` provenant du package `AICcmodavg`
- Le critère d'information Akaike (AIC) est une **mesure de qualité du modèle** pouvant être utilisée pour comparer les modèles
- `AICc` corrige pour le biais créé par les faibles tailles d'échantillon

# Coder et sélectionner le meilleur modèle

Pour trouver la valeur AICc d'un modèle, utilisez :

```
library(AICcmodavg)  
AICc(M1)  
# [1] 77.305
```

# Coder et sélectionner le meilleur modèle

Pour regrouper toutes les valeurs d'AICc dans un seul tableau, utilisez :

```
AICc ← c(AICc(M0), AICc(M1), AICc(M2), AICc(M3),  
         AICc(M4), AICc(M5), AICc(M6), AICc(M7), AICc(M8))  
  
Model ← c("M0", "M1", "M2", "M3", "M4", "M5", "M6", "M7", "M8")  
  
AICtable ← data.frame(Model = Model, AICc = AICc)
```

# Coder et sélectionner le meilleur modèle

Que signifient ces valeurs d'AICc ?

AICtable

#	Model	AICc
# 1	M0	479.85909
# 2	M1	77.30500
# 3	M2	35.49087
# 4	M3	277.29450
# 5	M4	457.66010
# 6	M5	269.10754
# 7	M6	461.82795
# 8	M7	81.02391
# 9	M8	31.84704

Le modèle avec le plus petit AICc a le plus grand pouvoir prédictif

Souvent on considère que deux modèles à +/- 2 unités d'AICc de différence ont une pouvoir prédictif équivalent

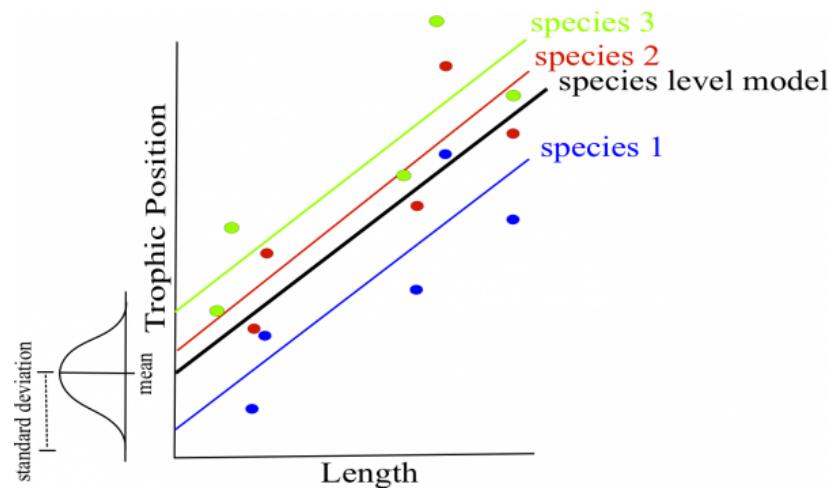
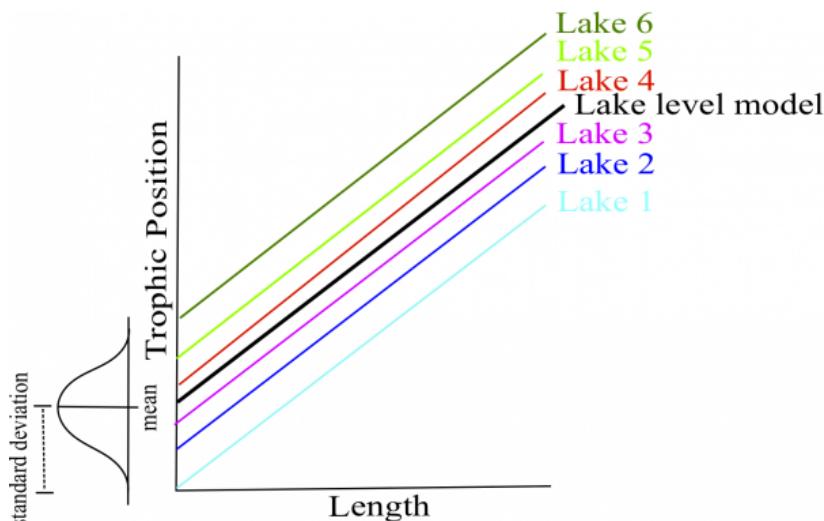
Regardons de plus près M8 et M2. On peut exclure les autres car ils ont des AICc tellement plus élevés

# Coder et sélectionner le meilleur modèle

Quelle est la structure du meilleur modèle?

```
M8 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 | Lake),  
           data = data, REML = FALSE)
```

L'intercept et l'effet de la longueur sur la position trophique peut varier selon l'espèce de poissons, mais seulement l'intercept peut varier par lac



# Coder et sélectionner le meilleur modèle

Une fois que les meilleurs modèles sont sélectionnés il faut remettre la méthode d'estimation à `REML=TRUE`

```
M8 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 | Lake),  
          data = data, REML = TRUE)
```

```
M2 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 + Z_Lengt  
          data = data, REML = TRUE)
```

# Défi 5



Prenez 2 minutes avec votre voisin pour étudier la structure du modèle M2.

Comment diffère-t-il de M8 d'un point de vue écologique?

Pourquoi n'est il pas surprenant que sa valeur d'AICc était la deuxième meilleure?

```
M8 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 | Lake),  
           data = data, REML = TRUE)
```

```
M2 ← lmer(Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 + Z_Length |  
           data = data, REML = TRUE)
```

# Solution

## Discussion de groupe...

**M2** La position trophique est une fonction de la longueur. L'intercept et l'effet de la longueur sur la position trophique peuvent varier selon l'espèce de poissons et le lac.

- les facteurs intrinsèques des espèces et des lacs sont à la base de relations différentes entre la position trophique et la longueur (i.e. pentes et intercepts)

**M8** La position trophique est une fonction de la longueur. L'intercept et l'effet de la longueur sur la position trophique peut varier selon l'espèce de poissons, mais seulement l'intercept peut varier par lac.

- seulement les facteurs intrinsèques des espèces sont responsables des différentes relations (i.e. pentes) et en moyenne, les positions trophiques pourraient être supérieures ou inférieures d'un lac à l'autre (e.g. intercepts).

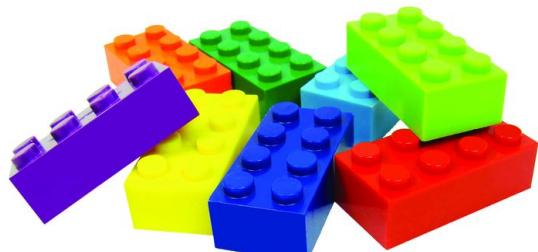
# Comment implémenter un MLM dans R?

Étape 1: Construction du modèle *a priori* et exploration des données

Étape 2: Coder les modèles potentiels et sélection du meilleur modèle

Étape 3: Validation du modèle

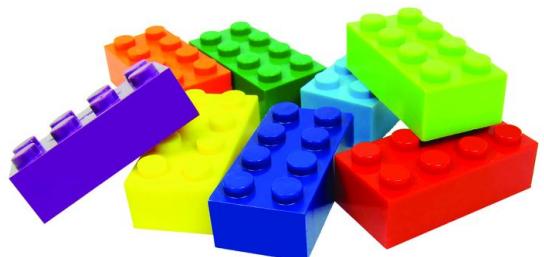
Étape 4: Interprétation et visualisation des résultats



# Étape 3 - validation du modèle

Vous devez vérifier que le modèle respecte toutes les suppositions de base:

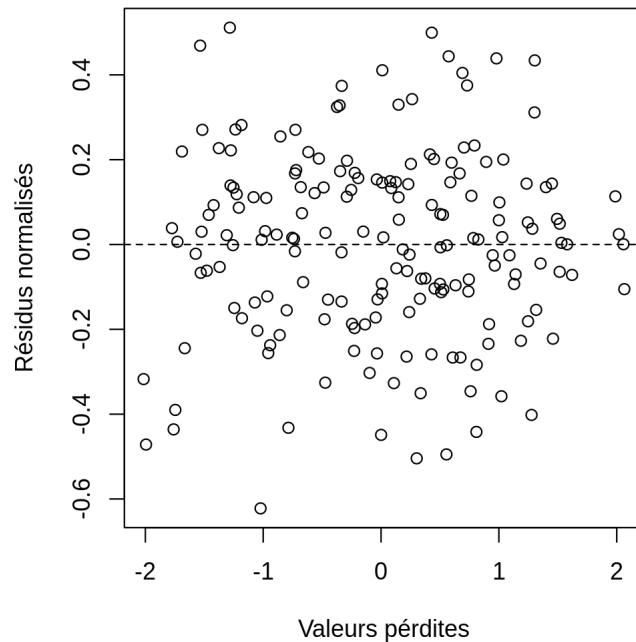
1. Vérifier l'homogénéité de la variance
  - Faire un graphique des valeurs prédictes en fonction des valeurs résiduelles
2. Vérifier l'indépendance des résidus
  - Graphique des résidus VS chaque covariable du modèle
  - Graphique des résidus VS chaque covariable non incluse du modèle
1. Vérifier la normalité
  - Histogramme



# Étape 3 - validation du modèle

1- Vérifier l'homogénéité de la variance

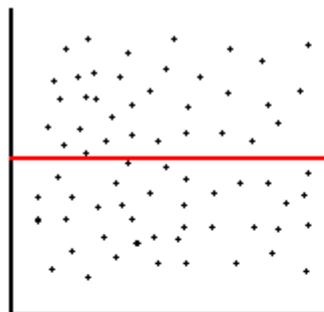
```
plot(resid(M8) ~ fitted(M8), xlab = 'Valeurs pérdites', ylab = 'Résidus n  
abline(h = 0, lty = 2)
```



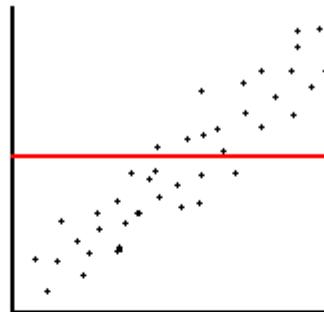
Étendue homogène des résidus → la supposition est respectée!

# Étape 3 - validation du modèle

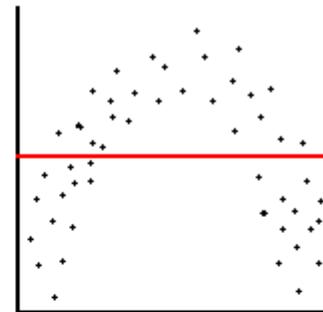
1- Vérifier l'homogénéité de la variance



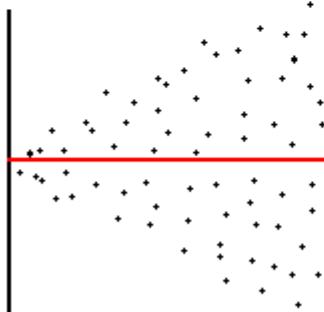
(a) Unbiased and Homoscedastic



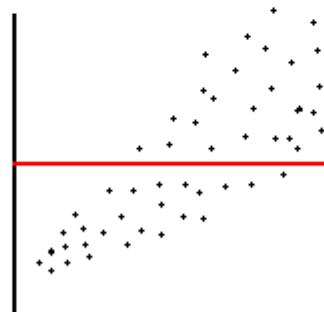
(b) Biased and Homoscedastic



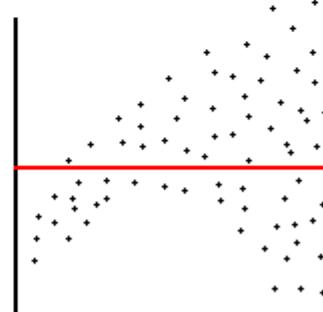
(c) Biased and Homoscedastic



(d) Unbiased and Heteroscedastic



(e) Biased and Heteroscedastic



(f) Biased and Heteroscedastic

# Étape 3 - validation du modèle

2- Vérifier l'indépendance des résidus avec chaque covariable

```
par(mfrow = c(1,3), mar=c(4,4,.5,.5))
```

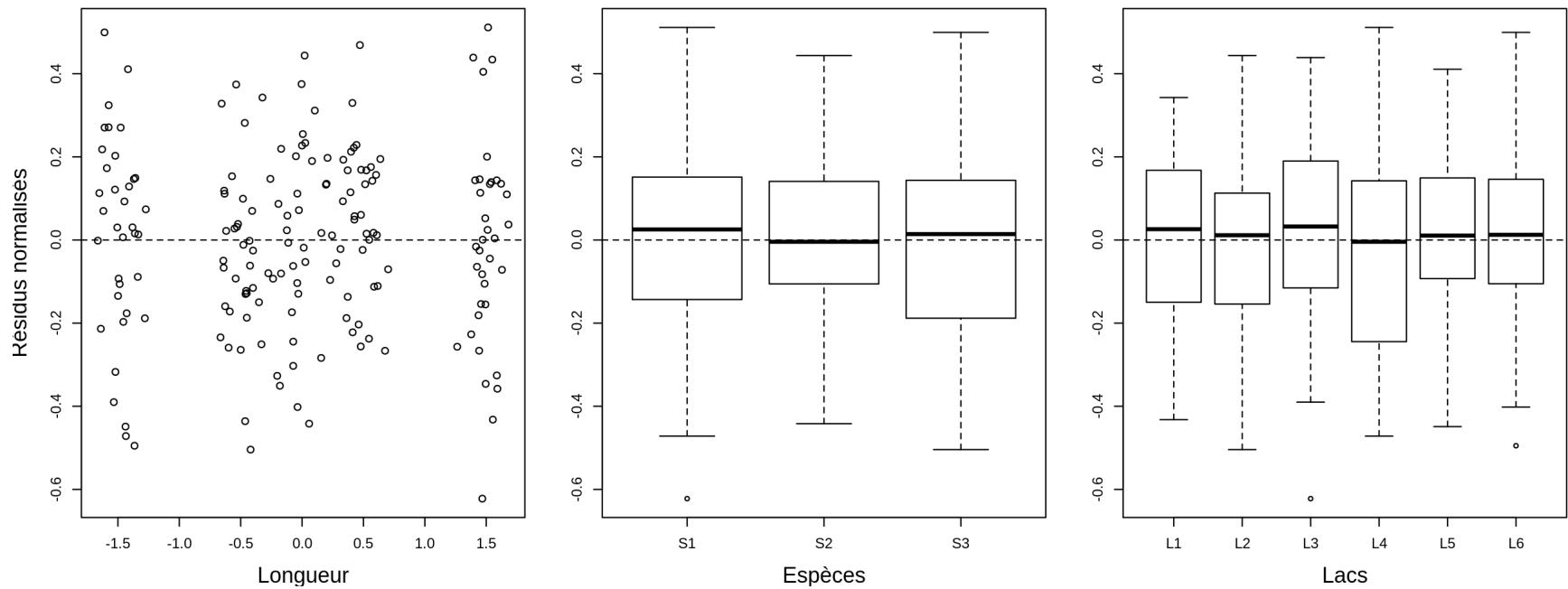
```
plot(resid(M8) ~ data$Z_Length, xlab = "Longueur", ylab = "Résidus normalisés")
abline(h = 0, lty = 2)
```

```
boxplot(resid(M8) ~ Fish_Species, data = data, xlab = "Espèces", ylab = "Résidus normalisés")
abline(h = 0, lty = 2)
```

```
boxplot(resid(M8) ~ Lake, data = data, xlab = "Lacs", ylab = "Résidus normalisés")
abline(h = 0, lty = 2)
```

# Étape 3 - validation du modèle

2- Vérifier l'indépendance des résidus avec chaque covariable



Étendue homogène des résidus autour de 0 → pas de patron des résidus en fonction de la variable, la supposition est respectée!

Note: Les regroupements de données sont dus à la structure des données, où des poissons de seulement 5 classes de taille (grand, petit, et trois groupes entre les deux) étaient capturés.

# Étape 3 - validation du modèle

2- Vérifier l'indépendance des résidus avec chaque covariable

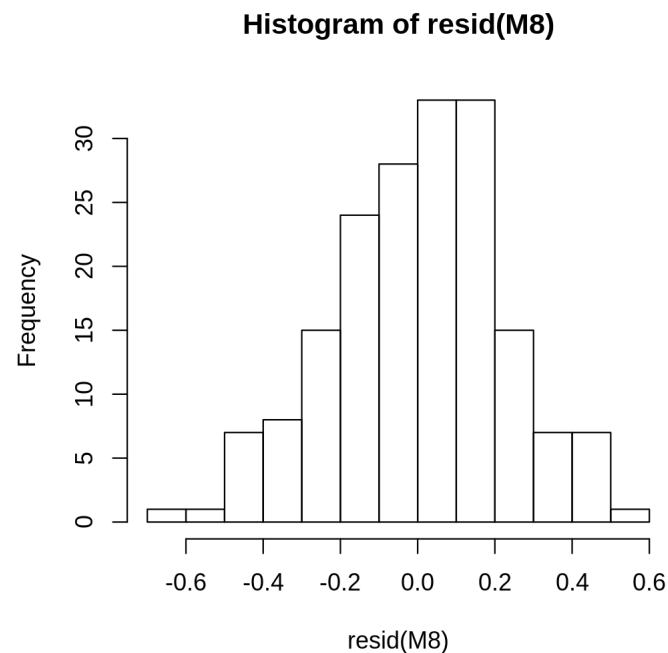
- Graphique des résidus VS chaque covariable non incluse du modèle
  - Si vous observez des patrons dans ce graphique, vous saurez qu'il y a de la variation dans votre jeu de données qui pourrait être expliquée par ces covariables. Vous devriez considérer d'inclure ces variables dans votre modèle.
  - Puisque dans notre cas, nous avons inclus toutes les variables mesurées dans notre modèle, nous ne pouvons pas faire cette étape.

# Étape 3 - validation du modèle

## 3- Vérifier la normalité des résidus

- Des résidus suivant une distribution normale indiquent que le modèle n'est pas biaisé

```
hist(resid(M8))
```



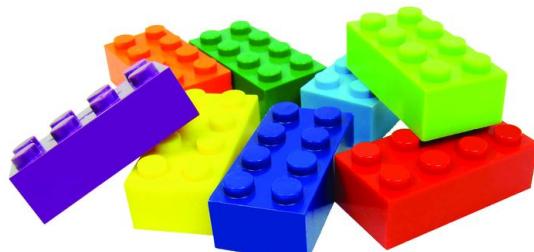
# Comment implémenter un MLM dans R ?

Étape 1: Construction du modèle *a priori* et exploration des données

Étape 2: Coder les modèles potentiels et sélection du meilleur modèle

Étape 3: Validation du modèle

Étape 4: Interprétation et visualisation des résultats



# Étape 4 - interprétation et visualisation

```
summ_M8 ← summary(M8)  
summ_M8
```

# Étape 4 - interprétation et visualisation

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Z_TP ~ Z_Length + (1 + Z_Length | Fish_Species) + (1 | Lake)
#   Data: data
#
# REML criterion at convergence: 21.7
#
# Scaled residuals:
#     Min      1Q  Median      3Q      Max
# -2.7718 -0.6016  0.0559  0.6424  2.2778
#
# Random effects:
# Groups           Name        Variance Std.Dev. Corr
# Lake            (Intercept) 0.20500  0.4528
# Fish_Species    (Intercept) 0.86621  0.9307
#                  Z_Length    0.02464  0.1570  1.00
# Residual
# Number of obs: 180, groups: Lake, 6; Fish_Species, 3
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept) -0.000906  0.568493 -0.002
```

# Étape 4 - interprétation et visualisation

```
# Random effects:  
# Groups           Name        Variance Std.Dev. Corr  
# Lake             (Intercept) 0.20500  0.4528  
# Fish_Species     (Intercept) 0.86621  0.9307  
#                 Z_Length     0.02464  0.1570   1.00  
# Residual
```

- **Groups**: facteurs groupant,
- **Name**:
  - **(Intercept)** pour l'ordonnée à l'origine,
  - ou le nom de la variable sur lequel porte l'effet mixe dans le cas d'une pente aléatoire, (**Z\_Length** dans notre exemple)
- **Variance** la variance estimée de l'effet (**Std.Dev.** est l'écart type de cette valeur)
- **Corr** indique la corrélation entre la pente aléatoire et l'ordonnée à l'origine aléatoire pour un groupement donné (voir **cette discussion** 

# Étape 4 - interprétation et visualisation

```
# Fixed effects:  
#  
#             Estimate Std. Error t value  
# (Intercept) -0.000906   0.568493 -0.002  
# Z_Length     0.422270   0.092170  4.581
```

Cette partie présente l'estimation des effet fixe. Une valeur de la statistique T (test de Student)](<https://en.wikipedia.org/wiki/T-statistic>) est retournée **sans p-value** (c'est un choix des auteurs du package, voir pourquoi dans [cette discussion](#)).

Cette statistique peut-être utilisée telle quelle. Vous pouvez aussi calculer l'intervalle de confiance (IC) à 95% avec cette table en utilisant

$$IC = Estimate \pm 1.96 \text{Std. Error}$$

Si 0 est dans cet interval, alors le paramètre n'est pas significativement différente de zéro au seuil  $\alpha = 0.05$ .

# Étape 4 - interprétation et visualisation

## Quelques fonctions utiles

- `coef(M8)` et `ranef(M8)` retourne les effets aléatoires du modèle M8
- `coef(summary(M8))` retourne les effets fixes
- `sigma(M8)` retourne l'écart type de la variance résiduelle
- `fitted(M8)` retourne les valeurs prédites par le modèle
- `residuals(M8)` retourne les résidus



## Défi 6

1. Quelle est la pente et son intervalle de confiance de la variable Z\_Length dans le modèle M8?
2. Est-ce que la pente de Z\_Length est significativement différente de 0 ?



# Solution

1. Quelle est la pente et son intervalle de confiance de la variable Z\_Length dans le modèle M8?
  - pente = 0.422;
  - limite supérieure de l'IC =  $0.4223 + 0.09 \cdot 1.96 = 0.5987$
  - limite inférieure de l'IC =  $0.4223 - 0.09 \cdot 1.96 = 0.2459$
2. Est-ce que la pente de Z\_Length est significativement différente de 0 ?
  - Oui, car l'IC [0.2459, 0.5987] n'inclut pas 0



# Défi 7

- Il est possible de visualiser graphiquement les différentes intercepts et pentes du modèle pour mieux interpréter les résultats

Prenez 2 minutes pour réfléchir aux différentes façons pour représenter les résultats de M8.

*Indice: considérez les différents "niveaux" du modèle*



# Solution

- a) Figure avec toutes les données regroupées
- b) Figure par espèce
- c) Figure par lac



# Solution

Pour faire ces figures, il nous faut:

- Les coefficients du modèle complet qui sont dans le résumé du modèle

```
summ_M8$coefficients
#                                     Estimate Std. Error      t value
# (Intercept) -0.000905958  0.56849279 -0.001593614
# Z_Length     0.422269906  0.09216952  4.581448649
```

- Intercept =  $-9.0595797 \times 10^{-4}$
- Pente = **0.4222699**



# Solution

Pour faire ces figures, il nous faut:

- Les coefficients pour chaque niveau du modèle qu'on obtient avec la fonction `coef`

```
coef(M8)
# $Lake
#     (Intercept) Z_Length
# L1 -0.085983970 0.4222699
# L2  0.002205106 0.4222699
# L3 -0.301816042 0.4222699
# L4 -0.574039102 0.4222699
# L5  0.218649633 0.4222699
# L6  0.735548627 0.4222699
#
# $Fish_Species
#     (Intercept) Z_Length
# S1  -1.0752954 0.2410627
# S2   0.5597853 0.5168365
# S3   0.5127923 0.5089106
#
```



# Solution

a) Figure avec toutes les données regroupées

```
library(ggplot2)

# Thème ggplot simplifié
fig ← theme_bw() +
  theme(panel.grid.minor=element_blank(), panel.grid.major=element_blank(),
        panel.background=element_blank()) +
  theme(strip.background=element_blank(), strip.text.y = element_text())
  theme(legend.background=element_blank()) +
  theme(legend.key=element_blank()) +
  theme(panel.border = element_rect(colour = "black", fill=NA))

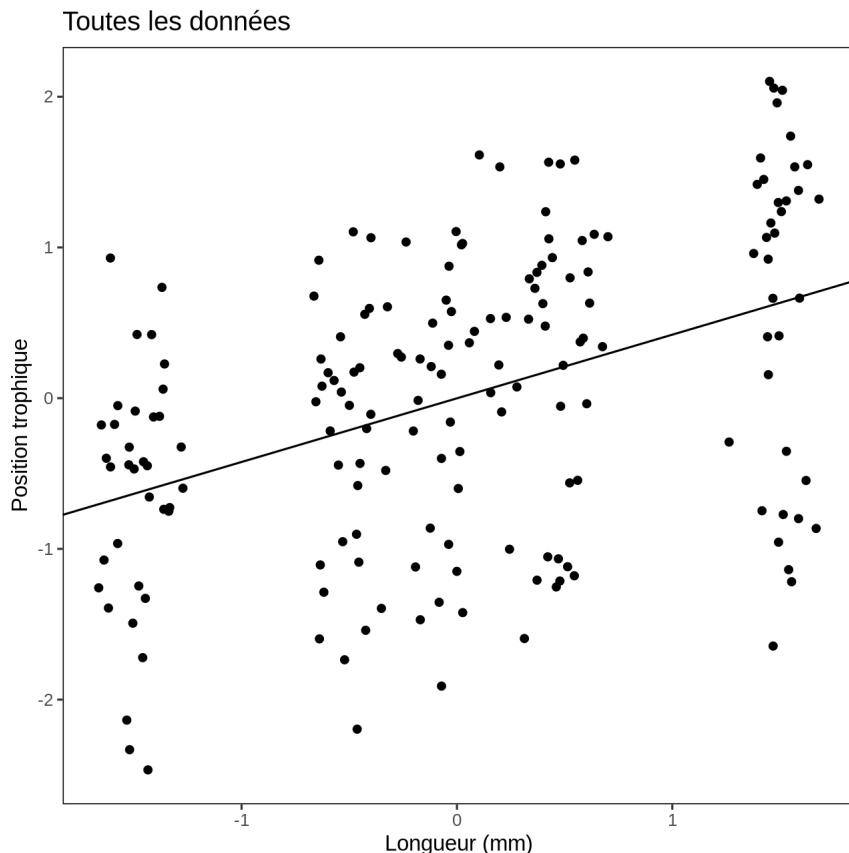
plot ← ggplot(aes(Z_Length, Z_TP), data = data)
Plot_AllData ← plot + geom_point() +
  xlab("Longueur (mm)") + ylab("Position trophique") +
  labs(title = "Toutes les données") + fig

Plot_AllData + geom_abline(intercept = -.0009059, slope = 0.4222697)
```



# Solution

a) Figure avec toutes les données regroupées





# Solution

b) Figure par espèce

```
# mettre les coefs dans un tableau pour les rendre plus faciles à manipuler
Lake.coef ← coef(M8)$Lake
colnames(Lake.coef) ← c("Intercept", "Slope")
Species.coef ← coef(M8)$Fish_Species
colnames(Species.coef) ← c("Intercept", "Slope")

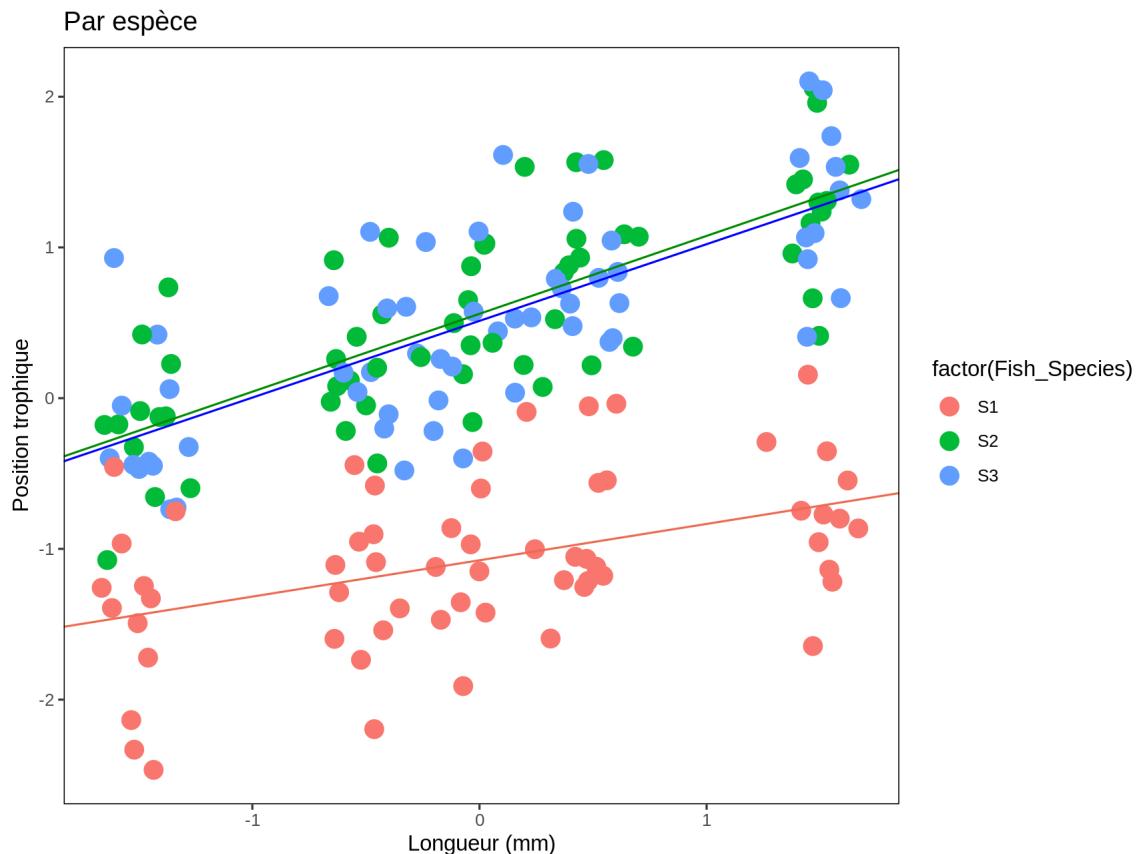
Plot_BySpecies←plot + geom_point(aes(colour = factor(Fish_Species)), size = 2)
xlab("Longueur (mm)") + ylab("Position trophique") +
  labs(title = "Par espèce") + fig

# Ajoutez les lignes de régression pour chaque espèce
Plot_BySpecies +
  geom_abline(intercept = Species.coef[1,1], slope = Species.coef[1,2], color = "red")
  geom_abline(intercept = Species.coef[2,1], slope = Species.coef[2,2], color = "blue")
  geom_abline(intercept = Species.coef[3,1], slope = Species.coef[3,2], color = "green")
```



# Solution

b) Figure par espèce





# Solution

c) Figure par lac

```
Plot_ByLake<-plot + geom_point(aes(colour = factor(Lake)), size = 4) +  
  xlab("Length (mm)") + ylab("Trophic Position") +  
  labs(title = "By Lake") + fig
```

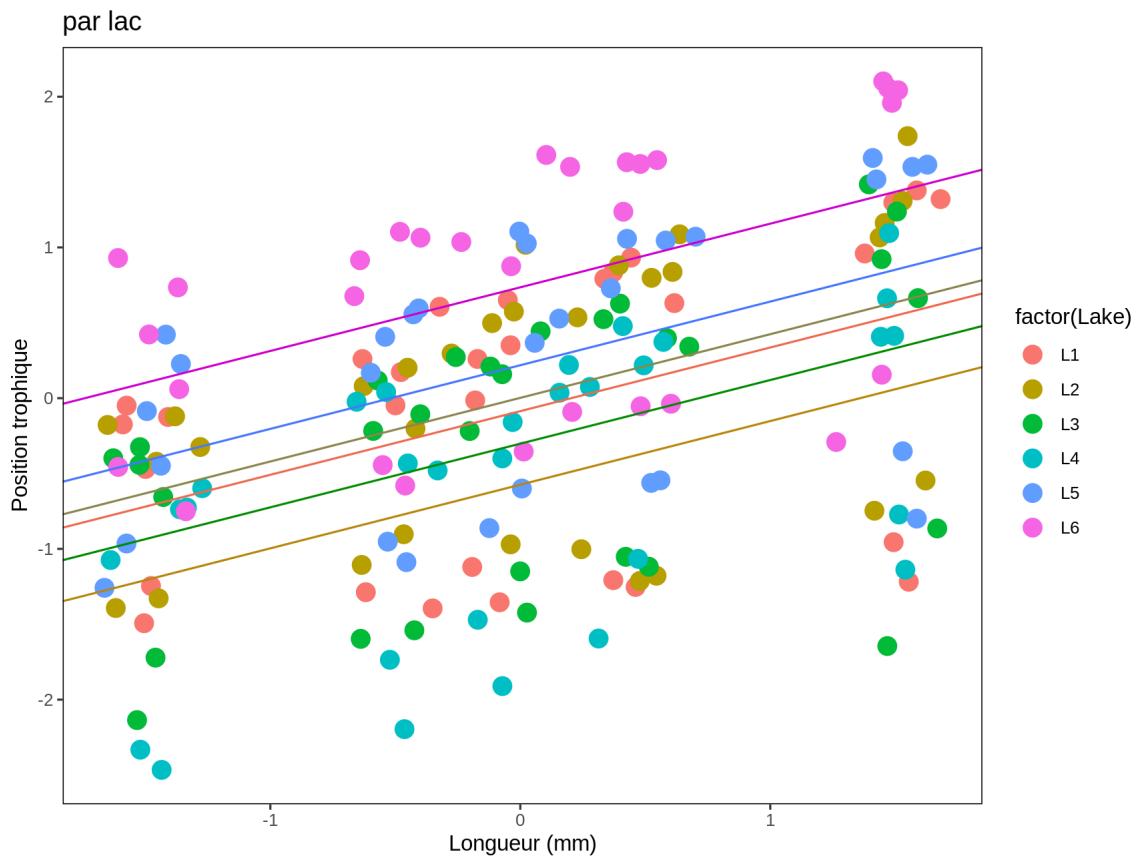
# Ajouter les lignes de régression avec les intercepts spécifiques à chaque lac

```
Plot_ByLake +  
  geom_abline(intercept = Lake.coef[1,1], slope = Lake.coef[1,2], col = "#0000FF")  
  geom_abline(intercept = Lake.coef[2,1], slope = Lake.coef[2,2], col = "#008000")  
  geom_abline(intercept = Lake.coef[3,1], slope = Lake.coef[3,2], col = "#800000")  
  geom_abline(intercept = Lake.coef[4,1], slope = Lake.coef[4,2], col = "#000080")  
  geom_abline(intercept = Lake.coef[5,1], slope = Lake.coef[5,2], col = "#808000")  
  geom_abline(intercept = Lake.coef[6,1], slope = Lake.coef[6,2], col = "#00FFFF")
```



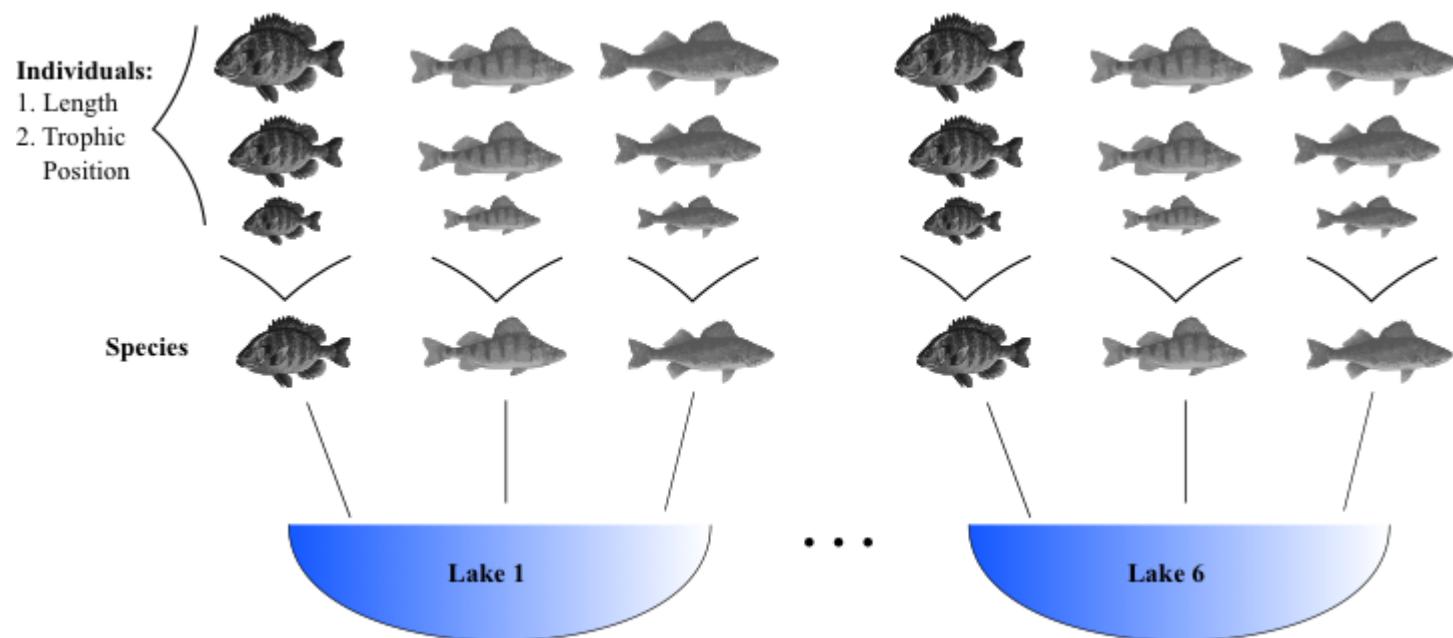
# Solution

c) Figure par lac



# Modèle mixtes et données en écologie

Les modèles mixtes sont très utiles pour prendre en compte la structure complexe des données en écologie tout en permettant de ne pas perdre beaucoup de degrés de liberté



# Défi 8



## Situation:

- Vous avez inventorié la richesse **dans 1000 quadrats** qui sont dans **10 sites différents** qui sont également dans **10 forêts différentes**.
- Vous avez de plus **mesuré la productivité** dans chaque **quadrat**.
- Vous désirez savoir si la productivité est un bon prédicteur de biodiversité

**Quel modèle mixte pourriez-vous utiliser pour ce jeu de données?**



# Solution!

```
lmer(Biodiv ~ Productivite + (1 | Foret / Site))
```

Ici les effets aléatoires sont nichés (i.e. Sites dans forêt) et non croisés.

Pourquoi utiliser `(1 | Foret / Site)` plutôt que `(1 | Foret) + (1 | Site)` ?

Regardez [cette réponse sur](#)  !

# Défi 9



## Situation:

- Vous avez récolté **200 poissons** dans **12 sites différents** distribués également dans **4 habitats** différents qui se retrouvent dans **un même lac**.
- Vous avez mesuré la **longueur de chaque poisson** et la **quantité de mercure dans ses tissus**.
- Vous désirez savoir si l'habitat est un bon prédicteur de la concentration en mercure.

**Quel modèle mixte pourriez-vous utiliser pour ce jeu de données?**



# Solution!

```
lmer(Mercure ~ Longueur * Habitat + (1 | Site))
```

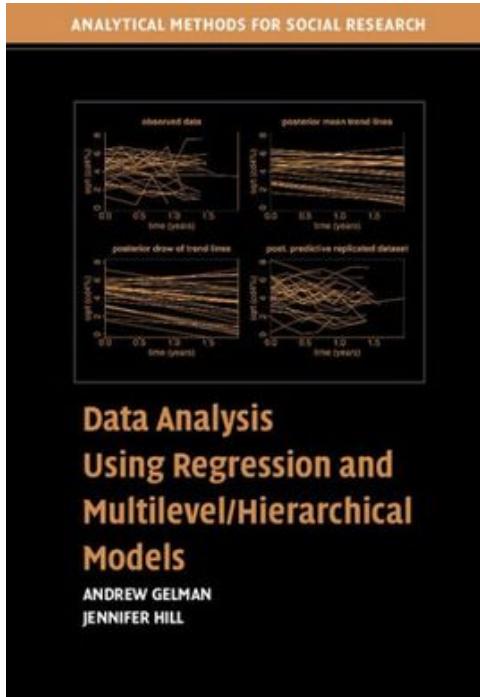
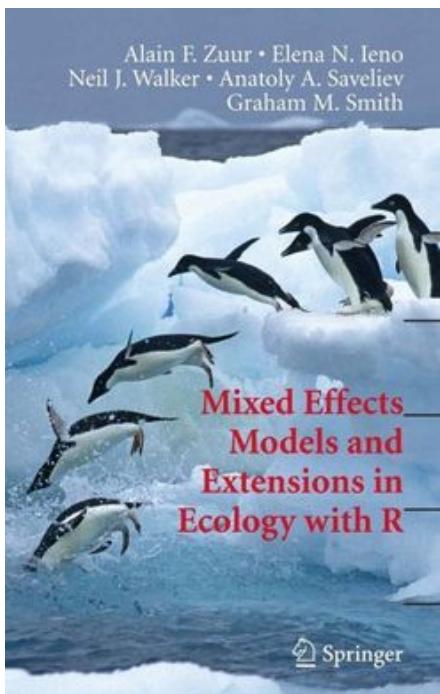
# Défi 10!



- Discutez du jeu de données sur lequel vous travaillez avec votre voisin et déterminez si un modèle mixte serait approprié.
- Si oui, travaillez ensemble pour écrire le code que vous utiliseriez pour faire ce modèle dans R.
- Si non, imaginez un jeu de données fictif pour lequel un modèle mixte serait approprié et codez ce modèle.

# Ressources additionnelles

- Différences entre `nlme` et `lme4`



- Harrison et al. (2018), PeerJ, DOI [10.7717/peerj.4794](https://doi.org/10.7717/peerj.4794)

**Merci de votre participation à cet atelier!**

