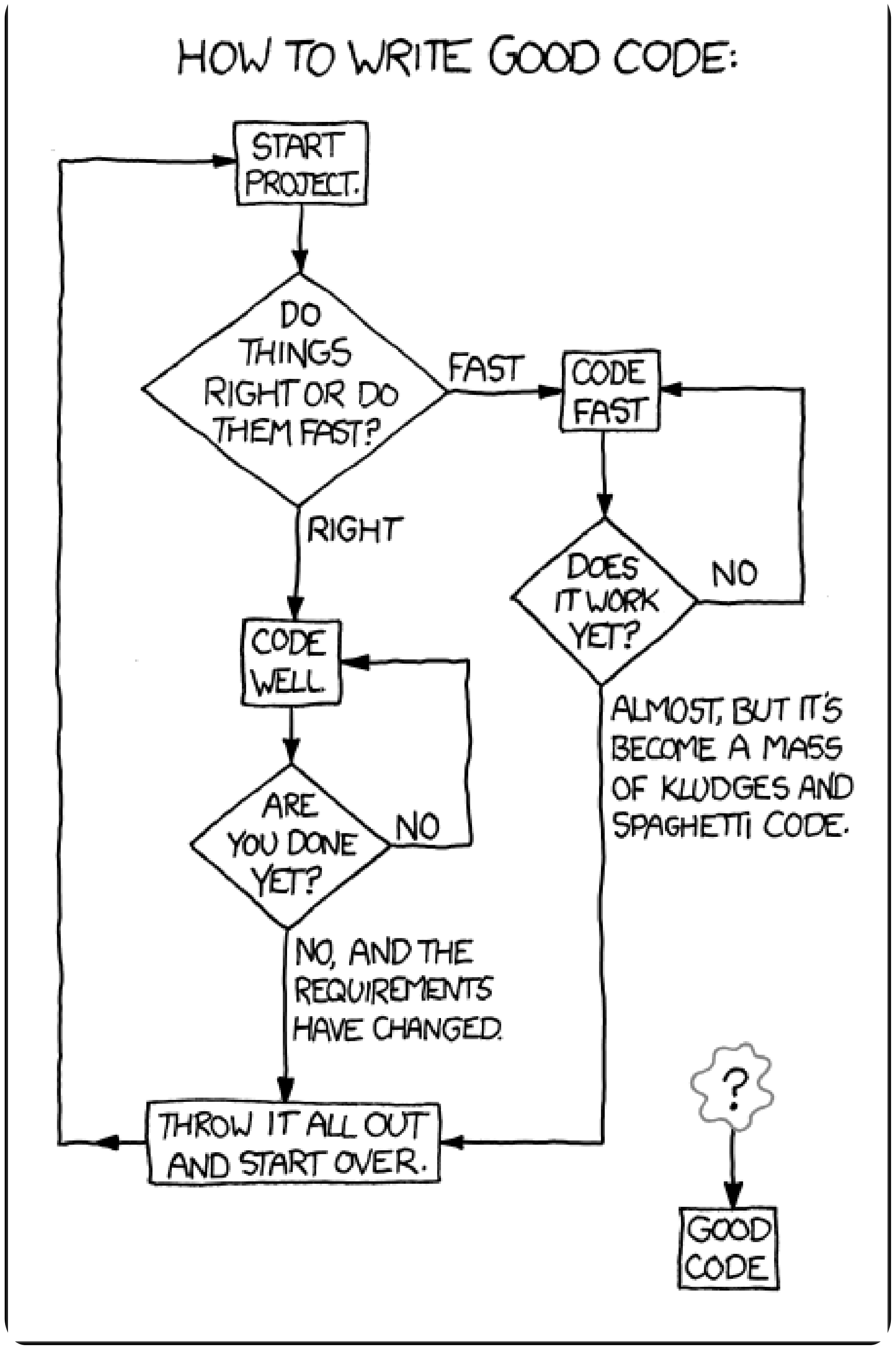

HOW TO WRITE GOOD CODE:
START PROJECT.
DO THINGS RIGHT OR DO THEM FAST?
FAST
CODE FAST
RIGHT
DOES IT WORK YET?
NO
CODE WELL
ARE YOU DONE YET?
NO
ALMOST, BUT IT'S BECOME A MASS OF KLUDGES AND SPAGHETTI CODE.
NO, AND THE REQUIREMENTS HAVE CHANGED.
THROW IT ALL OUT AND START OVER.
?
GOOD CODE

focus on 1st 2 examples

Linear Mixed Effects Models

Bio 599

Reminders

Thursday: LMM paper discussion (skip model selection section – we will cover this next week)

Check out the new schedule on Canvas

Office hours Monday 10:30-11:30, Thursday 4:00-5:00 or by appointment

Outline

- What is a LME?
- Random vs. fixed effects
- How do LMEs work?
- Assumptions & Fit
- Example

What is a LME model?

Lots of different names for essentially the same thing:

- a mixed-effects (or mixed model);
- a random-effects model;
- a hierarchical or multi-level model; or
- a random-intercept model.

It’s called mixed because it mixes both fixed and random effects

What is a LME model?

Because relationship b/w group can be similar & clustered. eg Same measurement on same species in a similar plot.

Mixed-effect models are useful for:

- repeated measures
- naturally clustered or hierarchical in nature
- the experimental or sampling design involves replication at multiple levels of hierarchy
- quantifying variability of a response across different levels of replication
- generalizing to a larger population of sample units
- Unequal sample sizes among groups

Random vs. Fixed Effects

What are fixed effects?

You want a fixed effect is you want to see if Species A differs from Species B. or Plot B or A

Predictor variables – if continuous must be fixed.

We are interested in comparing the effects of specific groups

If we did the study again, we would choose the same groups

Conclusions reached in the study about differences among groups can **only** be applied to the groups included in the study

Pinhiero and Bates 2000, Ch. 1

Fixed Effects

- No random effects
- Residuals are the **only** source of random variation
- lm(), gls(), glm(), gam() models are only fixed effects

residuals are your only source of variation

model1<- lm(**y ~ x**, data=mydata)

Fixed part of formula y~x

Pinhiero and Bates 2000, Ch. 1

Random effect can only be Categorical + not Continuous.

What are Random Effects?

Undetermined categories/categories of a variable that are NOT repeatable

Groups are only a subset of the realized possibilities drawn from a ‘global’ set of population

The individual groups are not of much interest, but their variance might be.

Used to model the correlation or dependency among observations within certain groups or clusters in the data

Allows us to make predictions for unmeasured groups

Random Effects

- Randomly sampled categories of a variable, representing groups or clusters of measurements or units.

add this if you want a random slope & Intercept.

```
model<-lme(y~x,random=~x|site, data=mydata)
```

Fixed

Random part of the formula: describes the random effects and grouping structure

For each of the descriptions below, determine if it is a fixed effect or random effect.

- Medical treatments in a clinical trial — fixed — treatment word
- Plants measured repeatedly — random, measurement within same plot tends to be similar than another plant.
- Replicate aquarium tanks random
- Levels of ocean acidification fixed — levels treatment
- Transect with multiple quadrats in a sampling survey random

For each of the descriptions below, determine if it is a fixed effect or random effect.

- Medical treatments in a clinical trial
- Plants measured repeatedly
- Replicate aquarium tanks
- Levels of ocean acidification
- Transect with multiple quadrats in a sampling survey

Another way to think about the difference between fixed vs random

Fixed effects:

- Yes, an experiment with the same treatment levels **could be repeated**

Random effects:

- Random effects could **not** be repeated exactly the same again

Why add random effects?

1. Make inferences for unmeasured groups.

- conclusions **can be generalized**

2. Efficient use of data

- Avoids pseudoreplication by controlling for spatial and temporal non-independence

2. Improves accuracy of parameter estimation

- Use data from all the groups to estimate the mean and variance of the global distribution of group means.

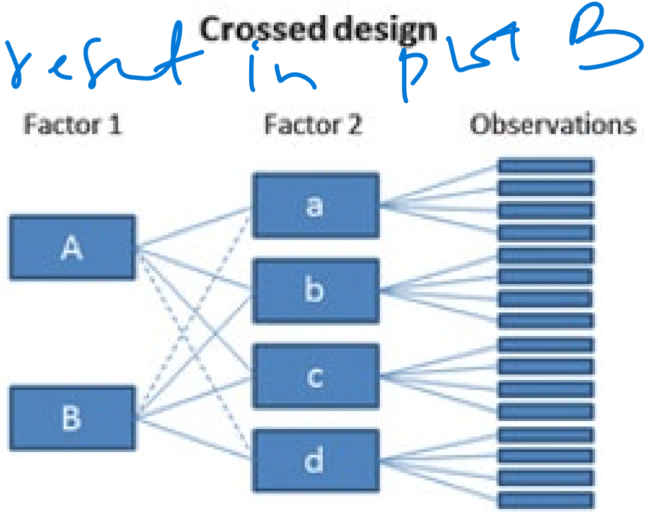
Pinhiero and Bates 2000, Ch. 1

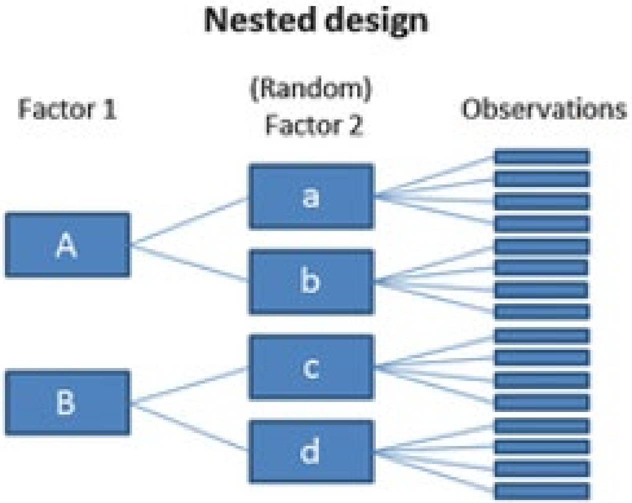
Crossed vs. Nested Random Effects

- Fixed and random factors can be nested or crossed
- Nested: some factor varies only within levels of another factor
- Crossed: the levels at which two factors vary are independent of each other

leaf from Species A in Plot A is also Present in plot B

nested leaf within tree
↳ within plot



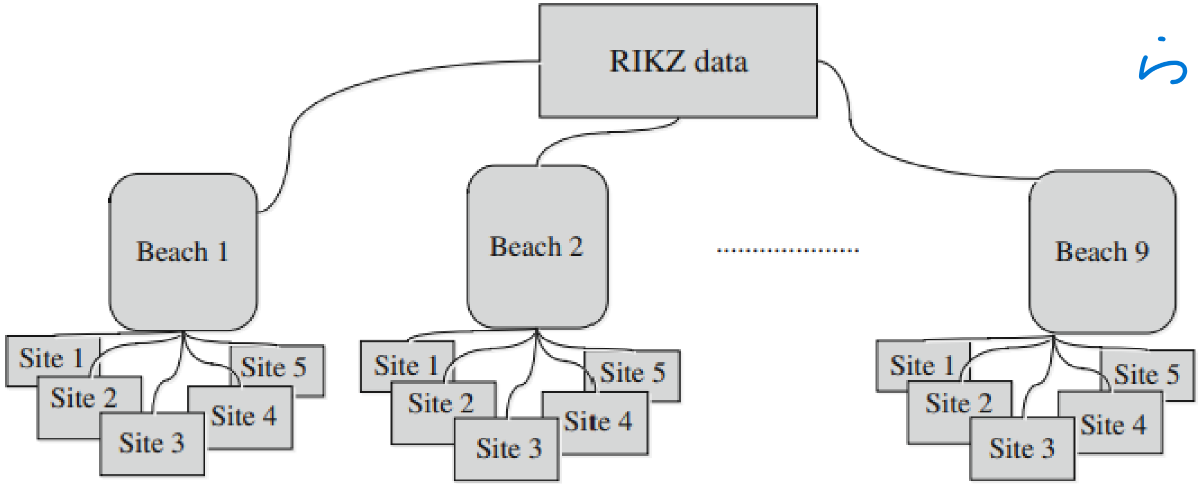


Schielzeth & Nakagawa, 2013 https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-210x.2012.00251.x

Do you have random effects? What are they?

- the way you specify your random effects will be determined by your experimental design
- visualize your experimental design by drawing it

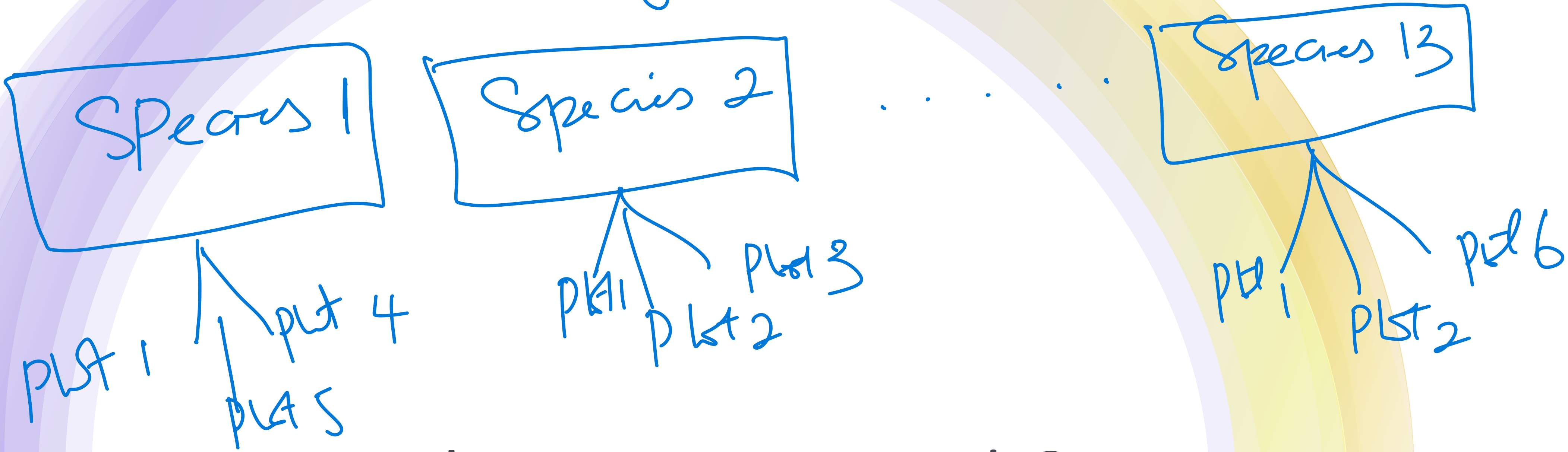
20 measurement per site, so site is nested within Beach.



– 1 measurement per site, so Beach is a random effect

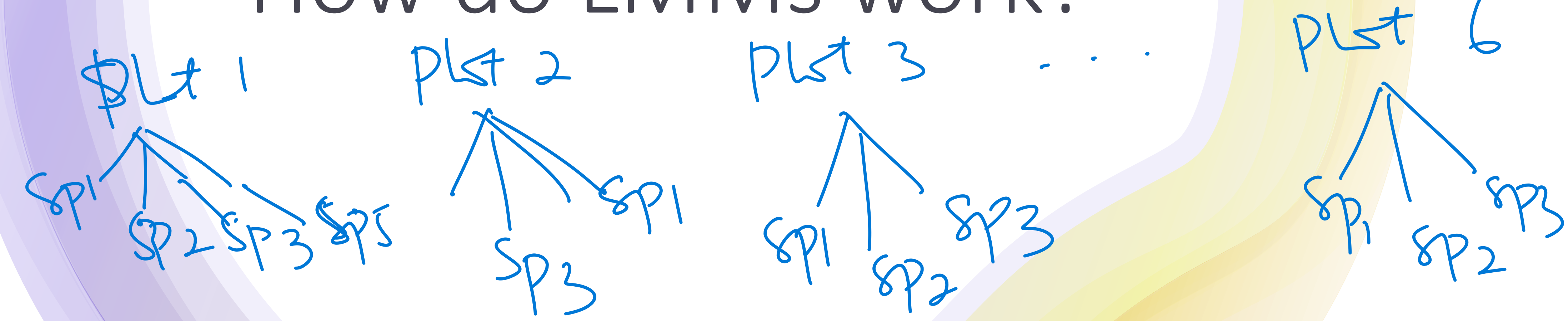
Fig. 5.1 Set up of the RIKZ data. Measurements were taken on 9 beaches, and on each beach 5 sites were sampled. Richness values at sites on the same beach are likely to be more similar to each other than to values from different beaches

Ogonna's data



How do LMMs work?

or



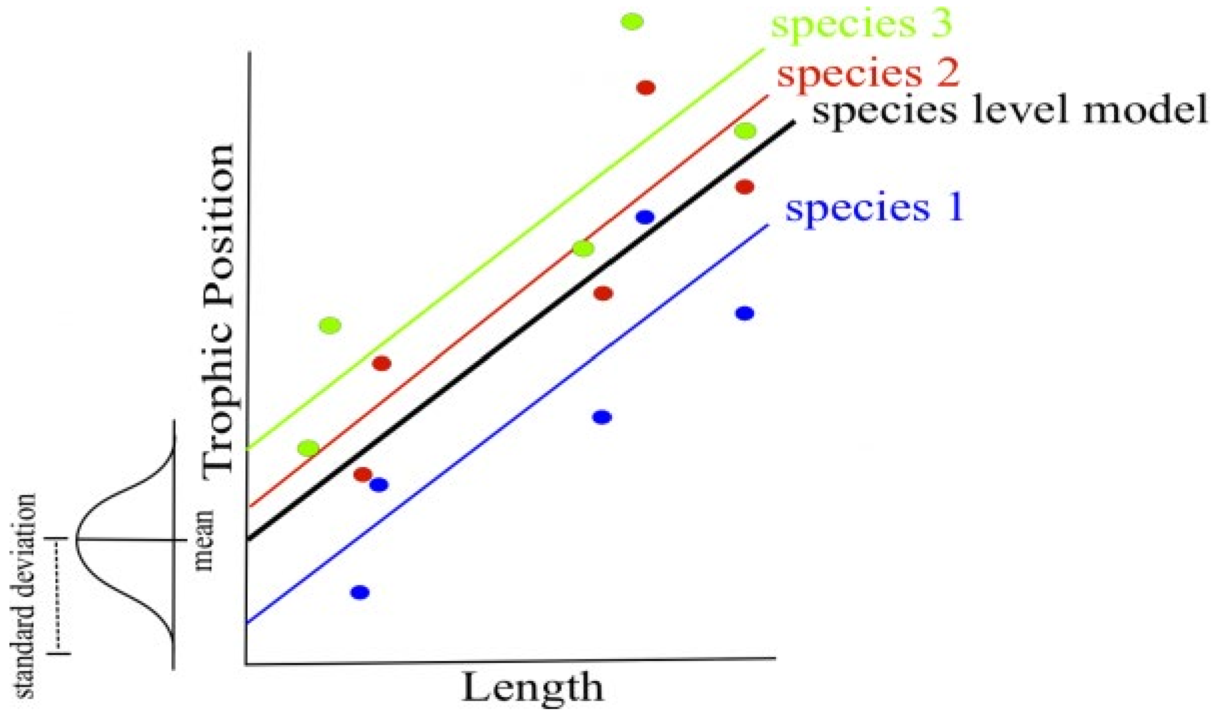
How do LMMs work?

1. Intercepts and/or slopes are allowed to vary according to a given factor, e.g. by lake and/or species.
   - Allowing intercepts and/or slopes to vary by random effects means that 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 (ML or REML).

2. Intercepts, slopes and their confidence intervals are adjusted to take the data structure into account.

Random Intercept

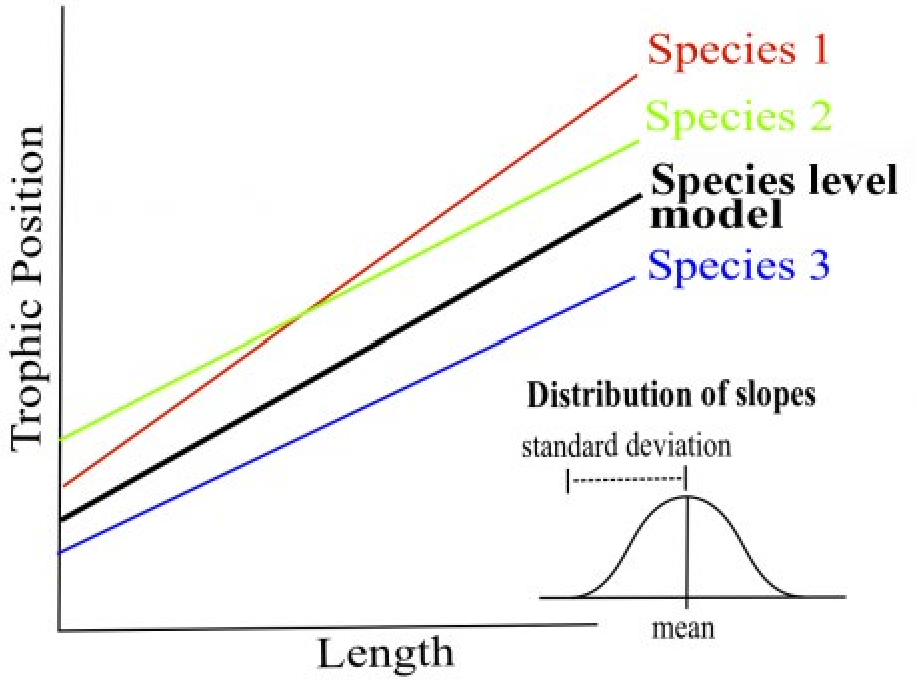
- It is assumed that the intercepts come from a normal distribution
- Only estimate the mean and standard deviation of the normal distribution instead of 3 intercepts, i.e. one for each species



- Note that the more levels your factor has, the more accurately the mean and standard deviation of the normal distribution will be estimated.
- Rule of thumb: 5 levels of the random grouping variable

Random Slope

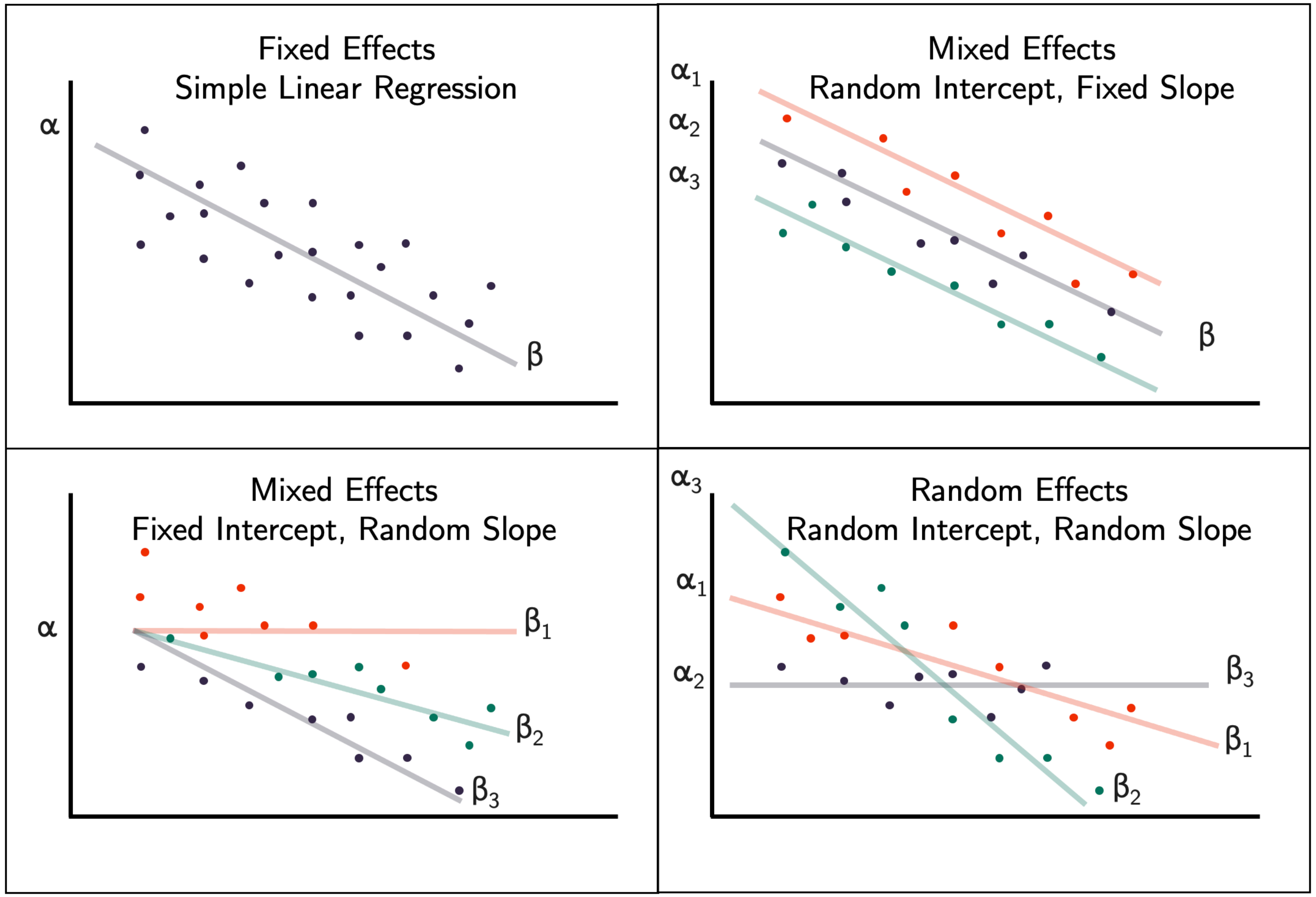
- The same principle applies to slopes that vary according to a given factor, only the mean and standard deviation of the slopes are estimated instead of three distinct slopes



random effects reduces rate of type1 & type 2
errors

Random Intercept and/or Slope

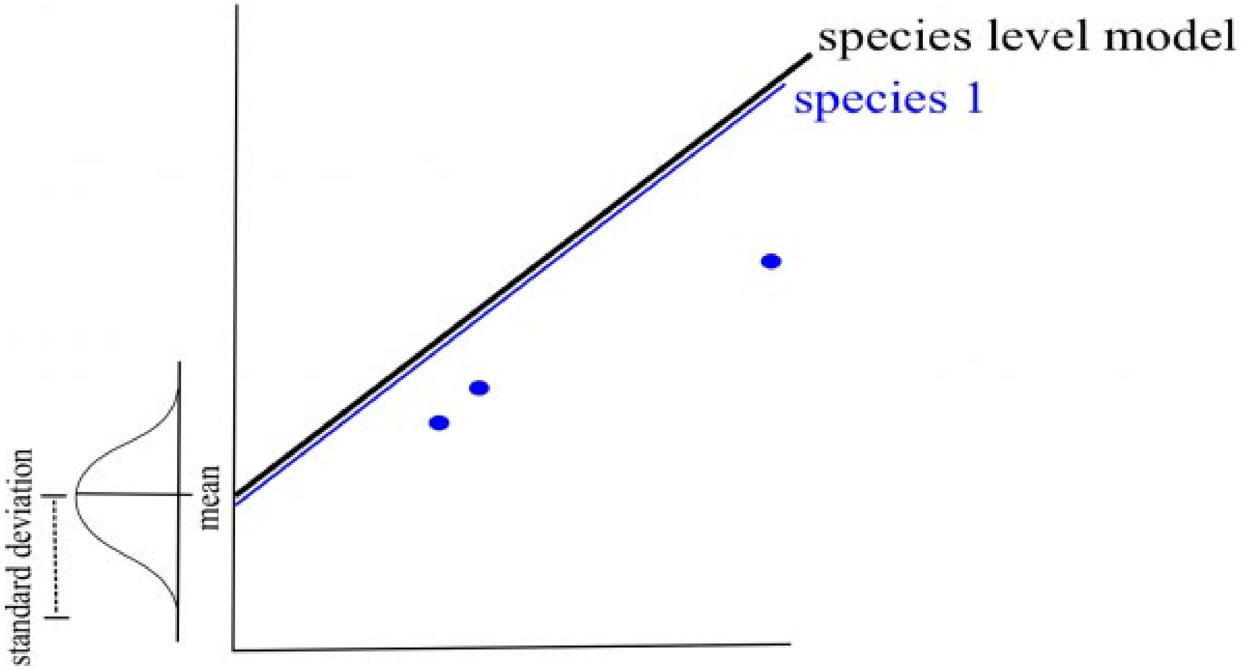
- Fitting random intercepts and slopes allows the slope of a predictor to vary based on a separate grouping variable
- When groups share a common slope =↑ error
- Always fit both random slopes and intercepts (*requires a lot of data)



Accounting for data structure

What happens if the sample size for a specific factor level is small?

- If a certain species is poorly represented (low *n*) in the data, the model will give more weight to the pooled model to estimate the intercept and slope of that species or lake = shrinkage
- Yes, LMEs can handle unequal and unbalanced sample sizes. BUT anova table calculations are approximate - use LRT.



Accounting for data structure

How do we assess the impact of a random effect on the model?

- The confidence intervals for the intercepts and slopes are adjusted to take account of the pseudo-replication-based on the **intraclass correlation coefficient (ICC)**
- How much variation is there in each VS group between groups?

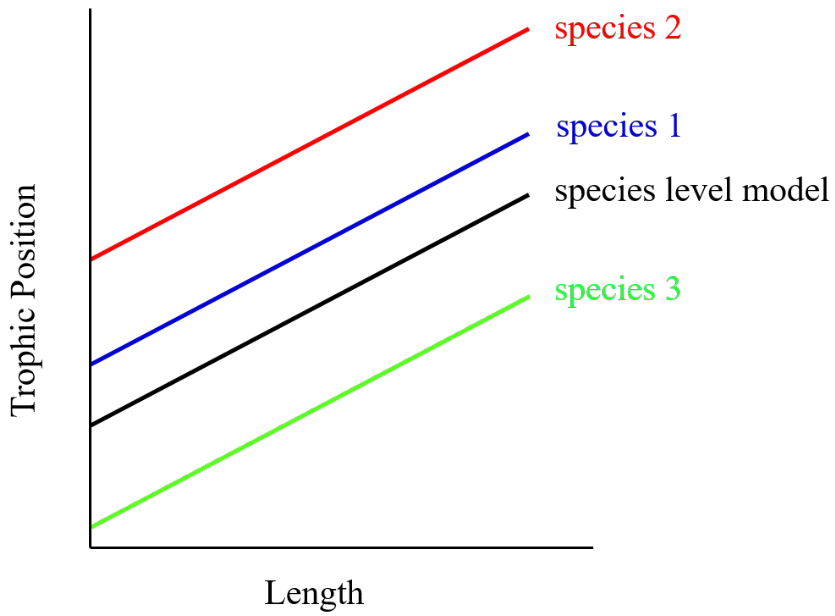
higher ICC, the higher the grouping structure matters

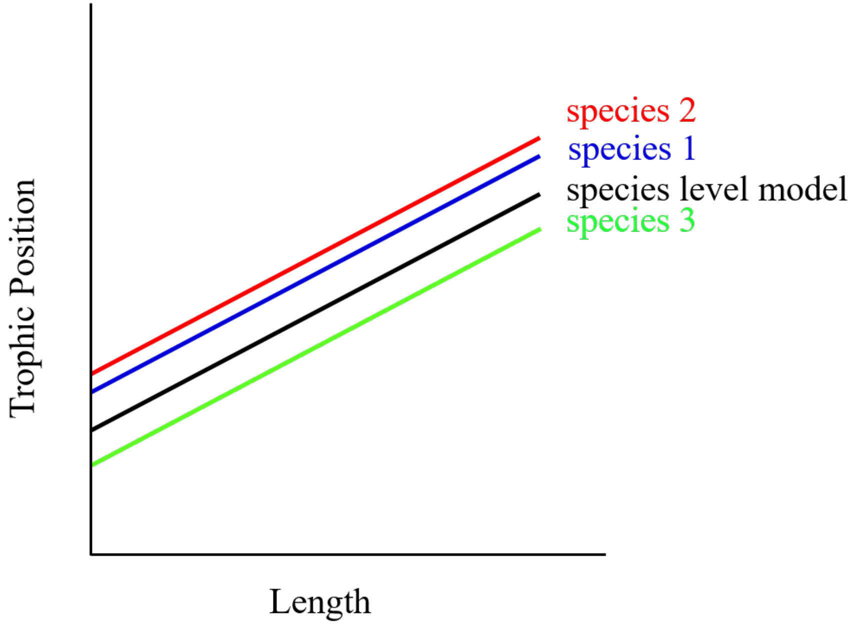
$$ICC = \frac{\boxed{\sigma_\alpha^2}}{\sigma_\alpha^2 + \boxed{\sigma_\varepsilon^2}}$$

between groups

within groups

Accounting for data structure





- ICC is high because **species differ** in their average trophic position.
- The confidence intervals for the general intercept is high.

- ICC is low because **species are similar** in their average trophic position.
- The confidence intervals for the general intercept is small.

LME Code

LME Code

Implemented with:

- nlme (Pinheiro et al., 2021)
    - can incorporate non-homogenous variance and autocorrelation
- lme4 (Bates et al., 2015)
    - can include crossed random effects

- Both packages model the variance structure of random effects explicitly.
- Differ in syntax and behind-the-scenes calculations but the **majority of basic LME theory is the same**

LME Code

lme (nlme package) vs lmer (lme4 package)

- You can use either for learning purposes and BIOL 599
- Example below is model with numerical X & Y variables

```
null.model<-lme(y~x,random=~x|animal, data=mydata)
null.model<-lmer(y~x+(x|animal),data=mydata)
```

Syntax for crossed and nested RE in lme4

- Intercept only model:

```
model<-lmer(y~x + (1|site), data=mydata)
```

- Slope only model:

```
model<-lmer(y~x + (0+x|site), data=mydata)
```

- Intercept & slope model:

```
model<-lmer(y~x + (1+x|site), data=mydata)
```

- Nested intercept model:

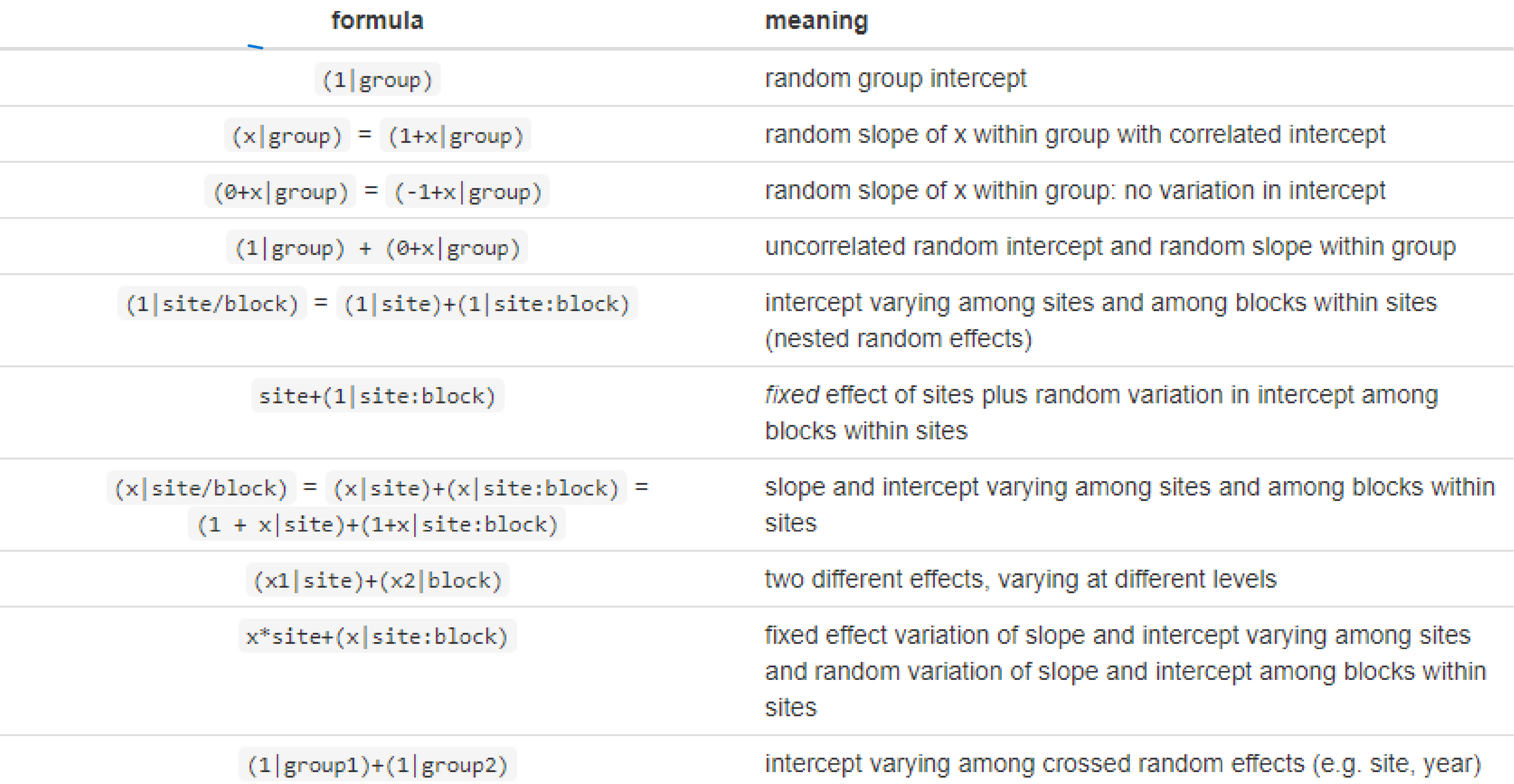
```
model<-lmer(y~x + (1|site/transect), data=mydata)
```

- Crossed intercept model:

```
model<-lmer(y~x + (1|site) + (1|year), data=mydata)
```

- Intercept, slope, crossed and nested:

```
model<-lmer(y~x + (1+x|site/transect) + (1|year), data=mydata)
```



formula	meaning
`(1|group)`	random group intercept
`(x|group)` = `(1+x|group)`	random slope of x within group with correlated intercept
`(0+x|group)` = `(-1+x|group)`	random slope of x within group: no variation in intercept
`(1|group) + (0+x|group)`	uncorrelated random intercept and random slope within group
`(1|site/block)` = `(1|site)+(1|site:block)`	intercept varying among sites and among blocks within sites (nested random effects)
`site+(1|site:block)`	*fixed* effect of sites plus random variation in intercept among blocks within sites
`(x|site/block)` = `(x|site)+(x|site:block)` = `(1 + x|site)+(1+x|site:block)`	slope and intercept varying among sites and among blocks within sites
`(x1|site)+(x2|block)`	two different effects, varying at different levels
`x*site+(x|site:block)`	fixed effect variation of slope and intercept varying among sites and random variation of slope and intercept among blocks within sites
`(1|group1)+(1|group2)`	intercept varying among crossed random effects (e.g. site, year)

Setting the likelihood: ML vs REML

- REML: restricted maximum likelihood is the *default*
  - lme(y~x, method="REML")
  - lmer(REML=TRUE)

use for random effect

- ML: maximum likelihood
  - lme (y~x, method="ML")
  - lmer(REML=FALSE)

use for fixed effect

REML for random

ML for fixed

REML for Final

Challenge #1

Situation:

- You have collected **200 fish** from **12 different sites** evenly distributed across **4 habitat types** that are found within **the same lake**.
- You measured **the length of each fish** and the **amount of mercury in its tissue**.
- You want to know if habitat and length are good predictors of mercury concentration.
- **What mixed model could you use for this dataset?**

site is a random effect here

Solution #1

```
lmer(Mercury ~ Length * Habitat_Type + (1 | Site))
```

Challenge #2

random

Site nested within forest

Productivity is the fixed effect.

Situation:

- You have inventoried species richness **in 1000 quadrats** that are within **10 different sites** which are also within **10 different forests**.
- You also **measured productivity** in each **quadrat**.
- You want to know if productivity is a good predictor of biodiversity
- **What mixed model could you use for this dataset?**

Solution #2

lmer(Biodiv ~ Productivity + (1 | Forest / Site))

- Here the random effects are nested (i.e. Sites within forest) and not crossed

Assumptions & Fit

Assumptions

- As with all linear models: Residuals follow a normal distribution with equal variance
- Groups are randomly sampled from a “population” of groups (i.e., are independent and sampled without bias).
- Replicates within groups are randomly sampled (independent)
- No carry-over between repeated measurements on the same subject.
- Within-group errors have constant variance
- Random effects (slopes and intercepts) have a normal distribution

Model Fit

What about R^2?

- **no direct equivalent of a traditional R^2 for LME models**
- LME models have variance associated with both random factor(variation between-groups) and residual variance of fixed factors (within-group variance)
- Difficult to compare R^2 from LME models among studies

Conditional and Marginal R^2

- fixed effects (marginal R2)
- random effects (conditional R2)

Nakagawa, Shinichi, and Holger Schielzeth. "A general and simple method for obtaining R2 from generalized linear mixed-effects models." *Methods in ecology and evolution* 4.2 (2013): 133-142.

LME example – Trophic Position and Size of Fish

Key Steps

Step 1: fit linear regression

Step 2: fit model with gls (so linear regression model can be compared with mixed-effects models)

Step 3: choose variance structure

- Introduce random effects, and/or
- Adjust variance structure to take care of heterogeneity

Step 4: fit the model

Step 5: compare new mixed-effects model with old

Step 6: validate model and, if necessary, repeat steps 4 and 5 until good model is found

Steps 7 and 8: find optimal fixed effects structure

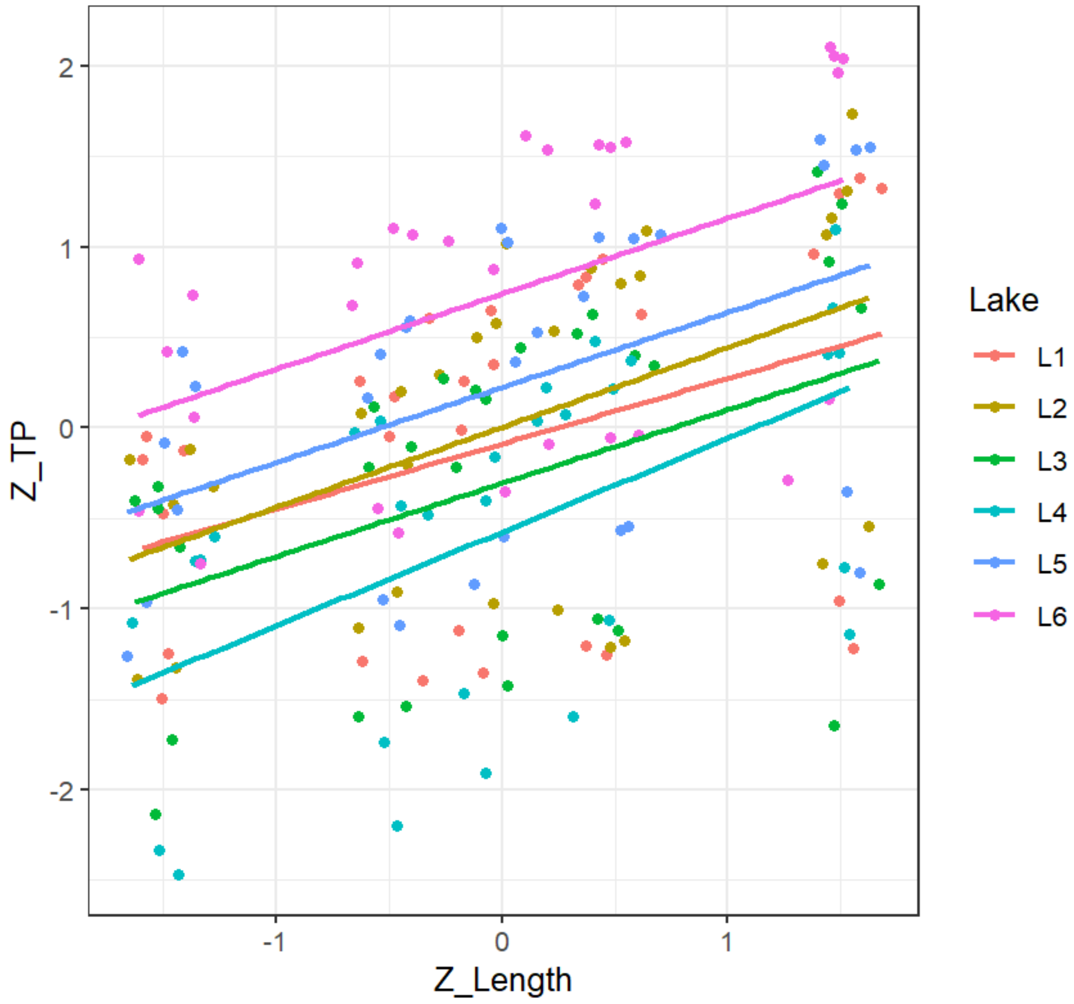
Step 9: refit with REML and validate model

Step 10: interpret

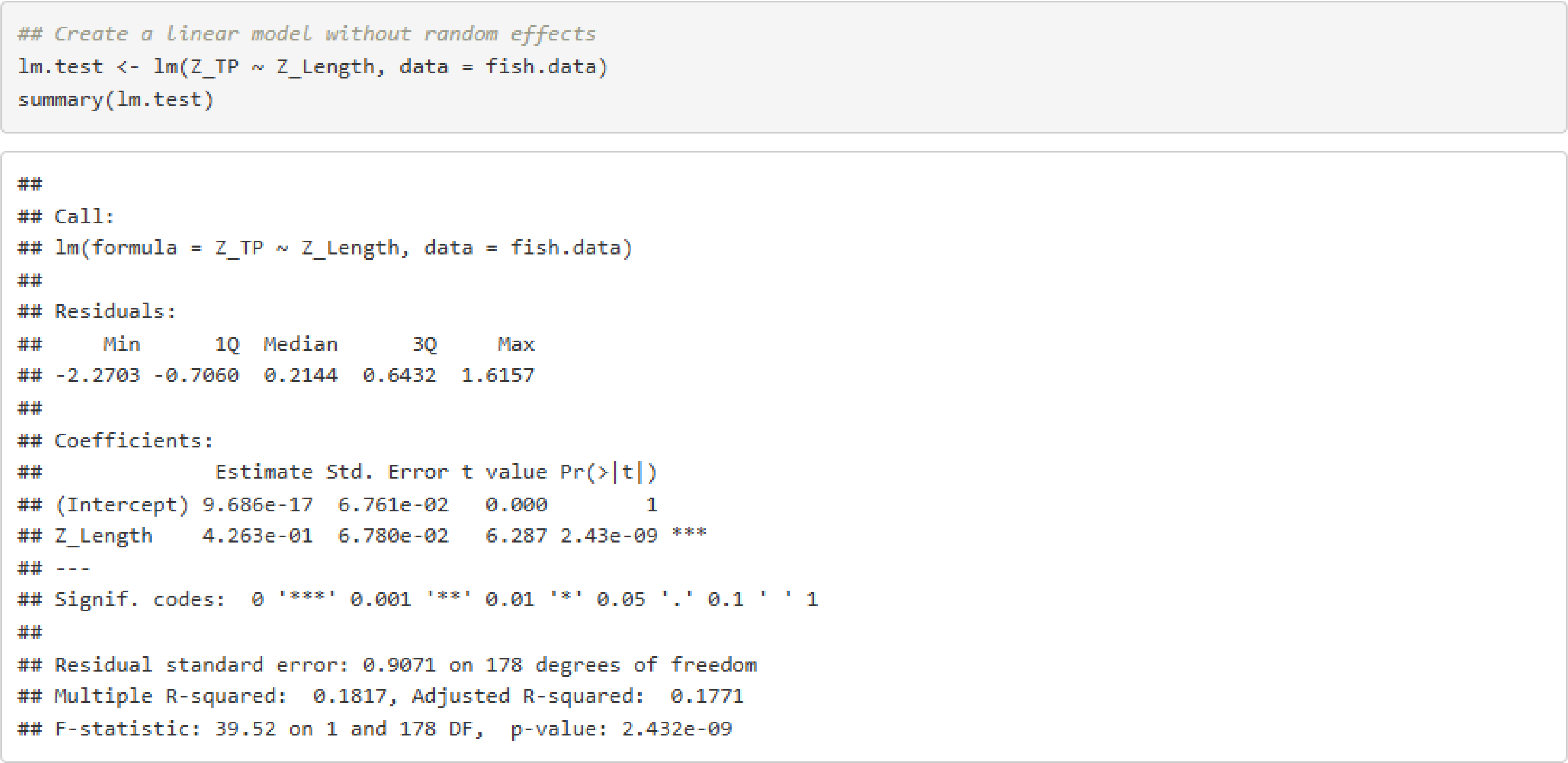
Example – fish data

Does fish trophic position increase with fish size?

- 30 fish per lake
- 6 different lakes
- 180 observations



Step 1: Linear Regression

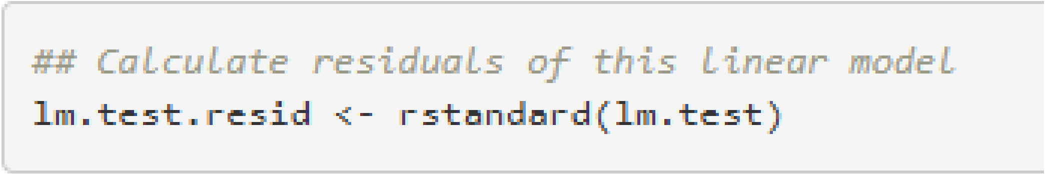


```
## Create a linear model without random effects
lm.test <- lm(Z_TP ~ Z_Length, data = fish.data)
summary(lm.test)
```

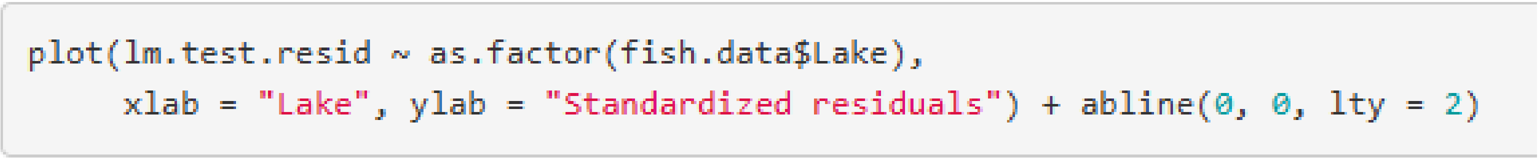
```
##
## Call:
## lm(formula = Z_TP ~ Z_Length, data = fish.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2703 -0.7060  0.2144  0.6432  1.6157
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.686e-17  6.761e-02   0.000        1
## Z_Length    4.263e-01  6.780e-02   6.287 2.43e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9071 on 178 degrees of freedom
## Multiple R-squared:  0.1817, Adjusted R-squared:  0.1771
## F-statistic: 39.52 on 1 and 178 DF,  p-value: 2.432e-09
```

NOTE - Step 2: GLS has same output

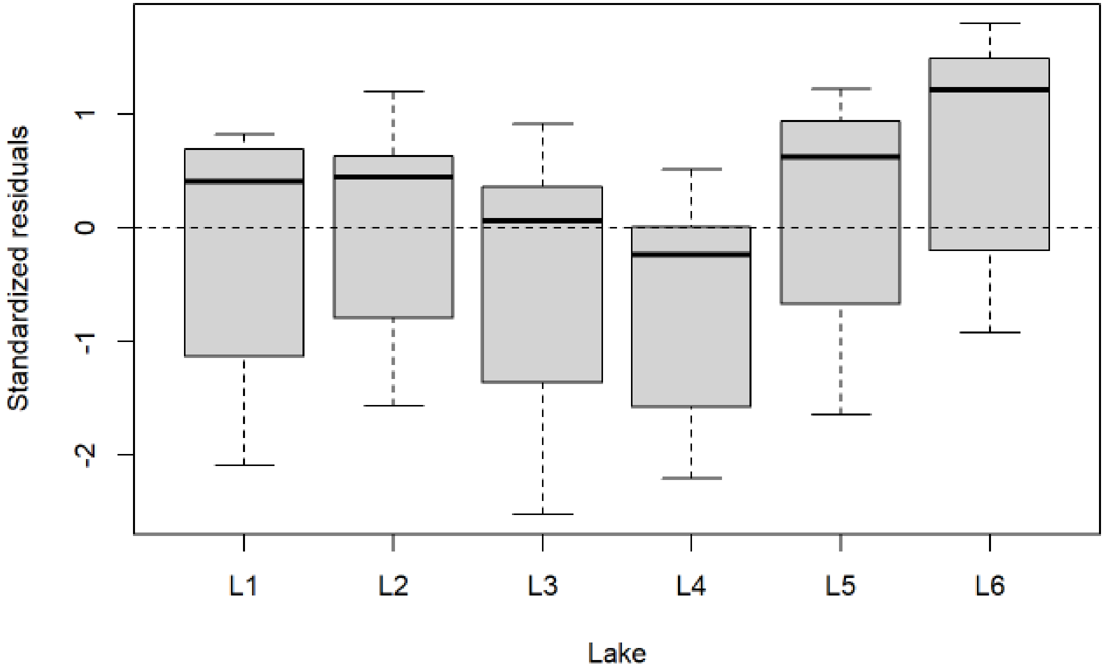
Step 3: Choose a variance structure



```
## Calculate residuals of this linear model
lm.test.resid <- rstandard(lm.test)
```



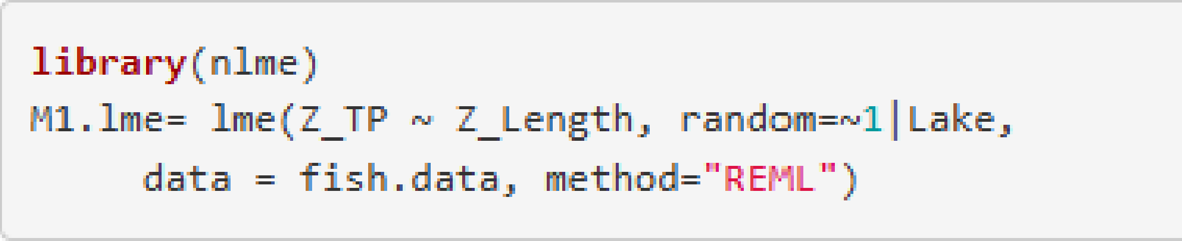
```
plot(lm.test.resid ~ as.factor(fish.data$Lake),
     xlab = "Lake", ylab = "Standardized residuals") + abline(0, 0, lty = 2)
```



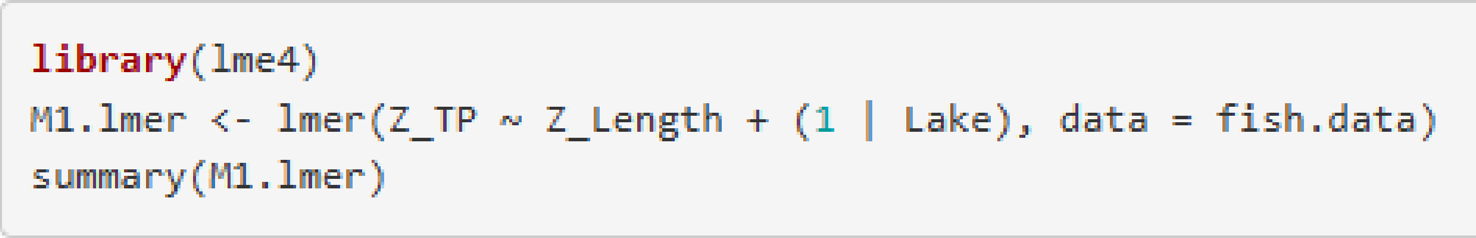
There is residual variance that could be explained by lake - we can account for this by including random intercepts for lake in the model

Include the random effect if it is part of your experimental design!!

Step 4: Fit the model



```r
library(nlme)
M1.lme= lme(Z_TP ~ Z_Length, random=~1|Lake,
     data = fish.data, method="REML")
```

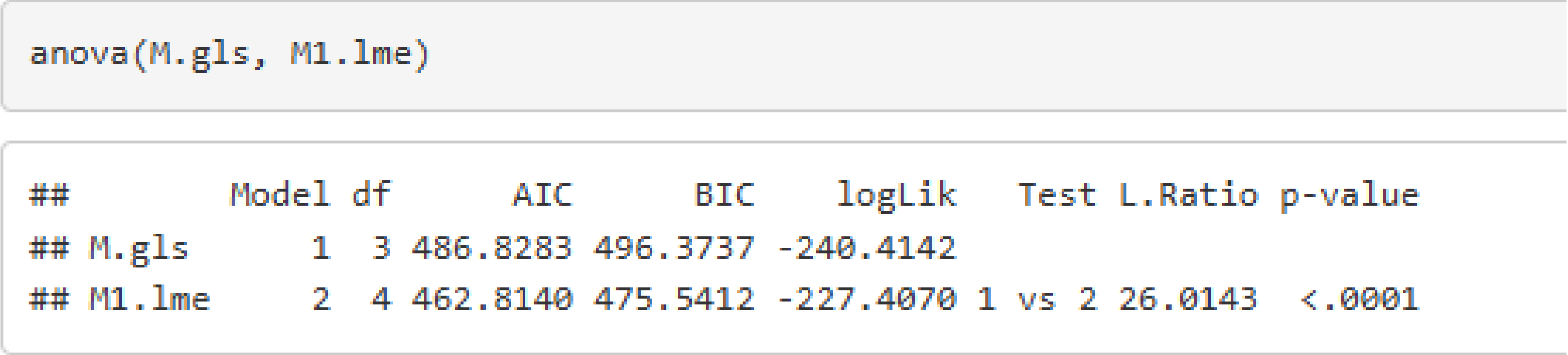


```r
library(lme4)
M1.lmer <- lmer(Z_TP ~ Z_Length + (1 | Lake), data = fish.data)
summary(M1.lmer)
```

lme() supports correlation structures and weighted variance, which can be useful in ecological data with heteroscedasticity or temporal autocorrelation

lmer() allows for more flexible random effect structures (e.g., random slopes), but does not directly support correlated residual structures or variance weights.

Step 5: Compare New and Old Models

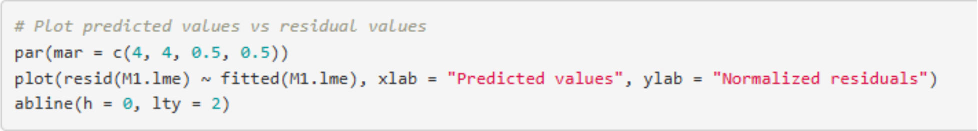


```
anova(M.gls, M1.lme)
```

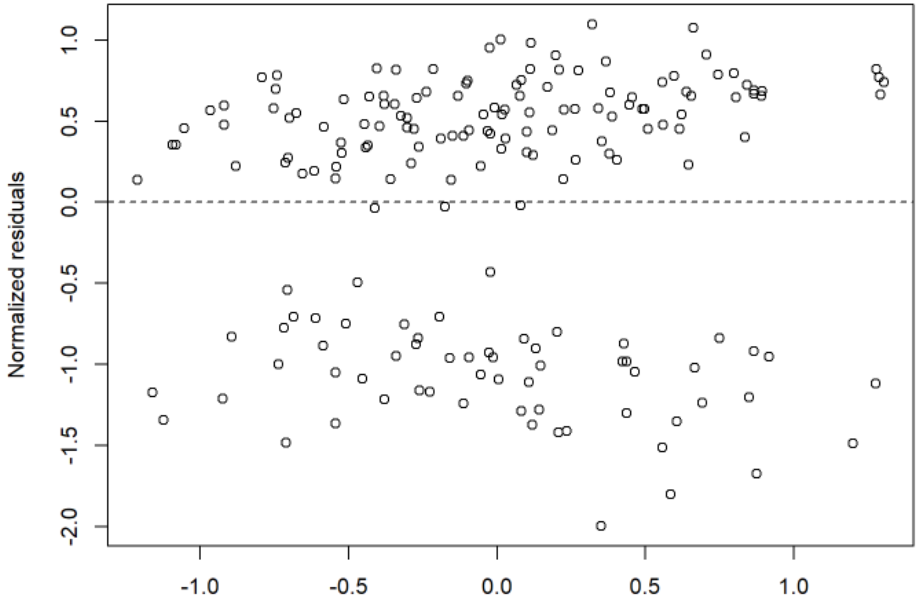
```
##        Model df      AIC      BIC    logLik   Test L.Ratio p-value
## M.gls      1  3 486.8283 496.3737 -240.4142
## M1.lme     2  4 462.8140 475.5412 -227.4070 1 vs 2 26.0143  <.0001
```

Step 6: Assumptions

Homogeneity of variance

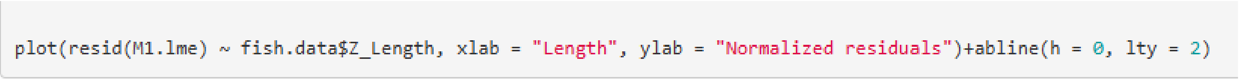


```
# Plot predicted values vs residual values
par(mar = c(4, 4, 0.5, 0.5))
plot(resid(M1.lme) ~ fitted(M1.lme), xlab = "Predicted values", ylab = "Normalized residuals")
abline(h = 0, lty = 2)
```

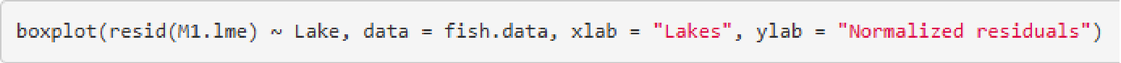


Step 6: Assumptions

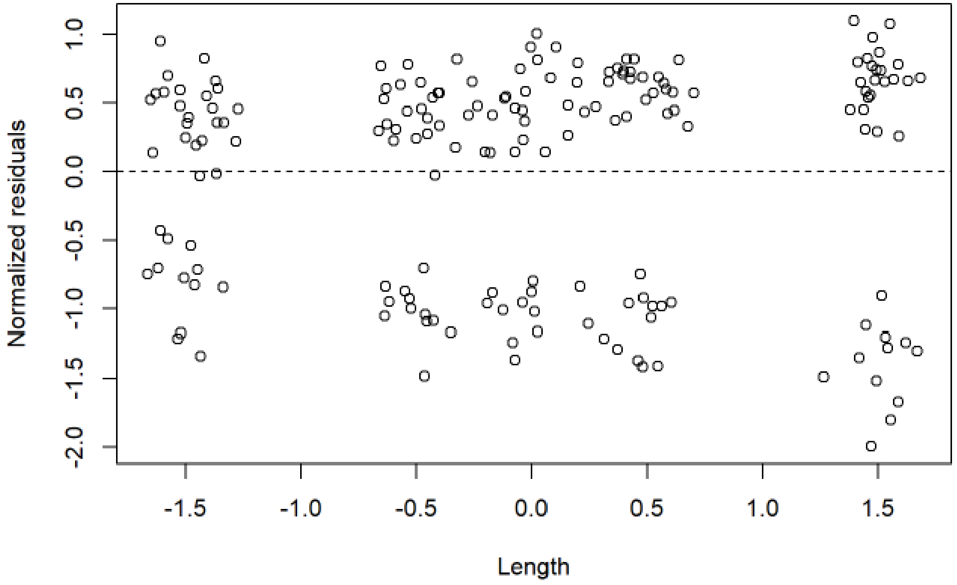
Independence

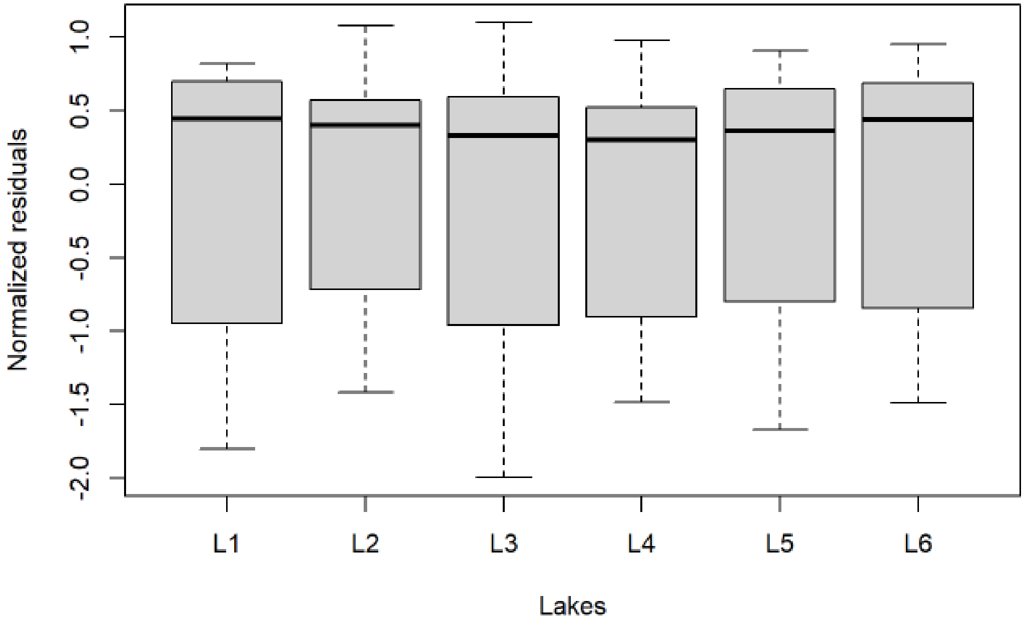


```
plot(resid(M1.lme) ~ fish.data$Z_Length, xlab = "Length", ylab = "Normalized residuals")+abline(h = 0, lty = 2)
```



```
boxplot(resid(M1.lme) ~ Lake, data = fish.data, xlab = "Lakes", ylab = "Normalized residuals")
```

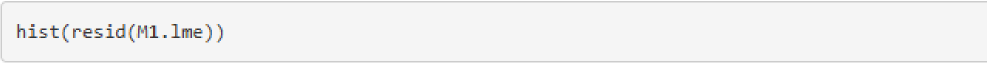




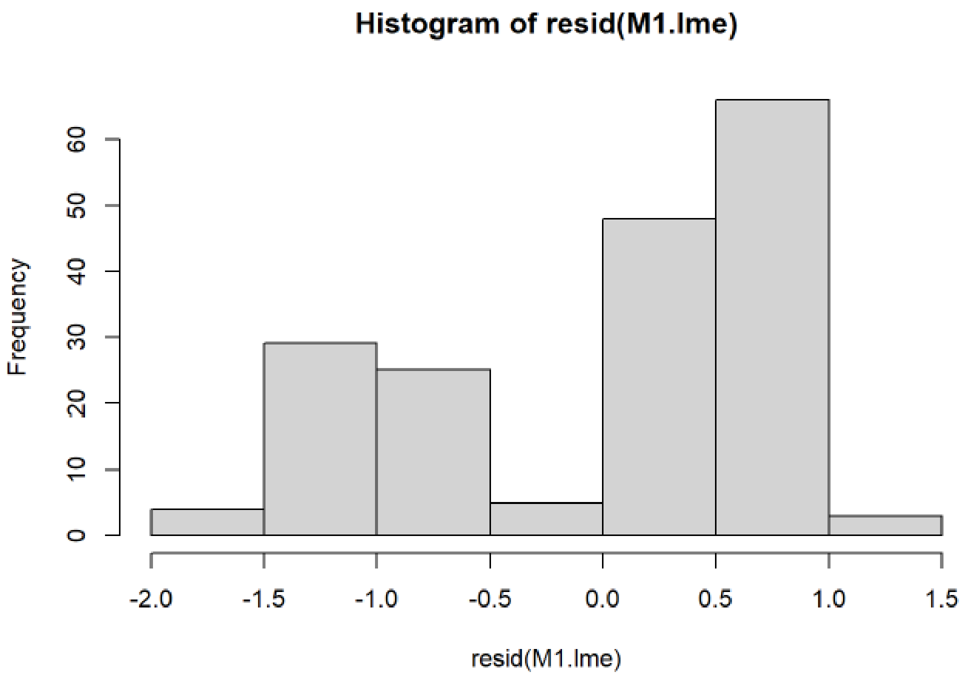
We should also plot residuals vs each covariate not included in the model

Step 6: Assumptions

- Normality of the model residuals

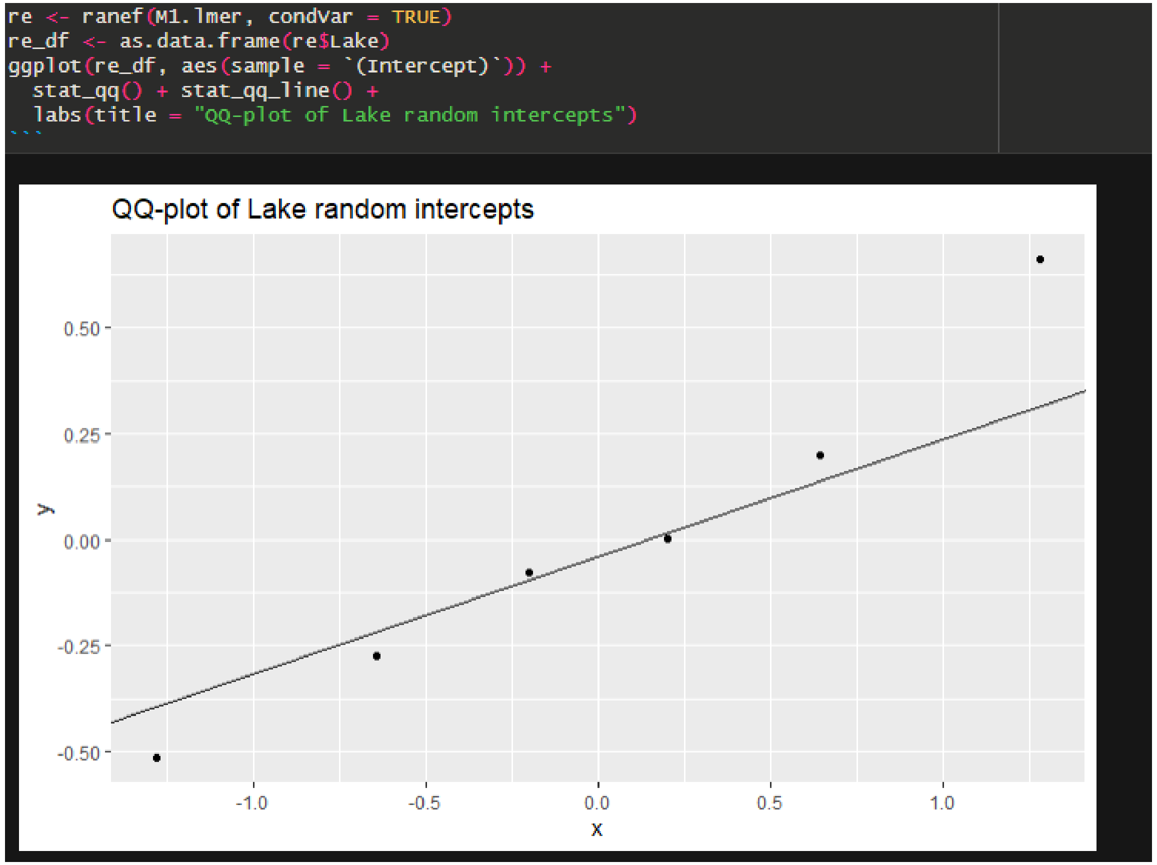


```
hist(resid(M1.lme))
```



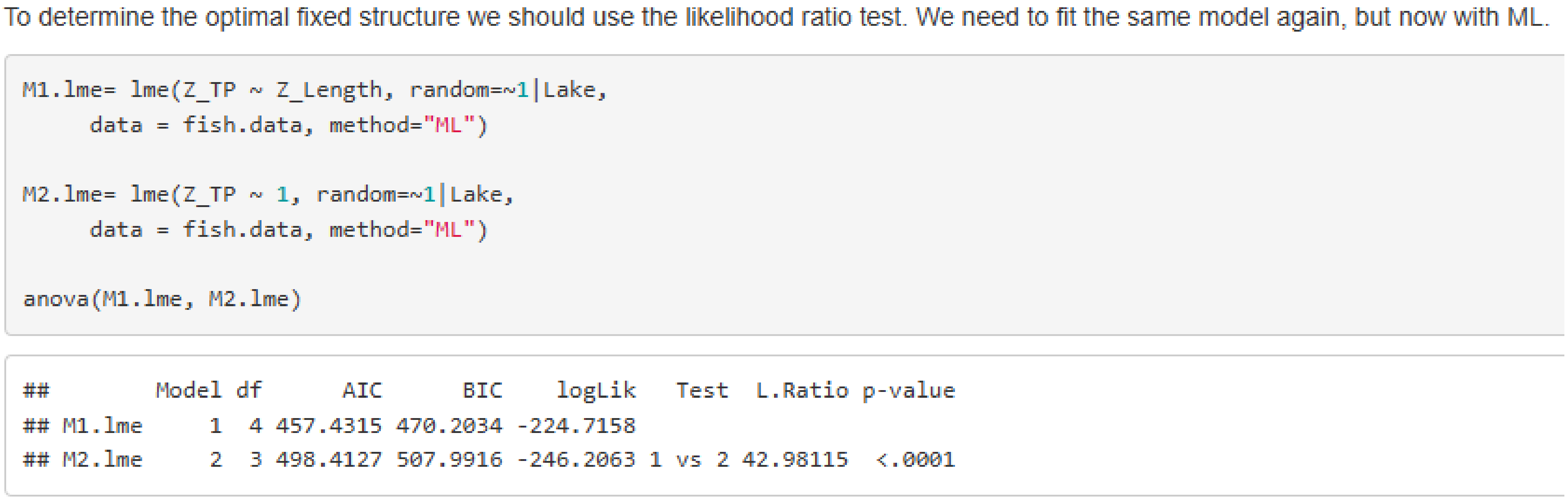
Step 6: Assumptions

- Normality of the random intercepts



~~~r
re <- ranef(M1.lmer, condVar = TRUE)
re_df <- as.data.frame(re$Lake)
ggplot(re_df, aes(sample = `(Intercept)`)) +
  stat_qq() + stat_qq_line() +
  labs(title = "QQ-plot of Lake random intercepts")
```
~~~

Step 7 and 8: Optimal fixed effects structure



To determine the optimal fixed structure we should use the likelihood ratio test. We need to fit the same model again, but now with ML.

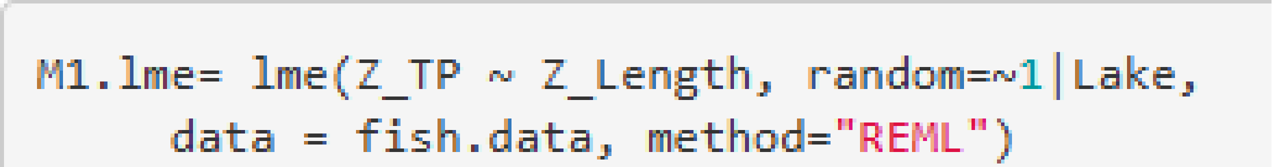
```
M1.lme= lme(Z_TP ~ Z_Length, random=~1|Lake,
     data = fish.data, method="ML")

M2.lme= lme(Z_TP ~ 1, random=~1|Lake,
     data = fish.data, method="ML")

anova(M1.lme, M2.lme)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## M1.lme     1  4 457.4315 470.2034 -224.7158
## M2.lme     2  3 498.4127 507.9916 -246.2063 1 vs 2 42.98115  <.0001
```

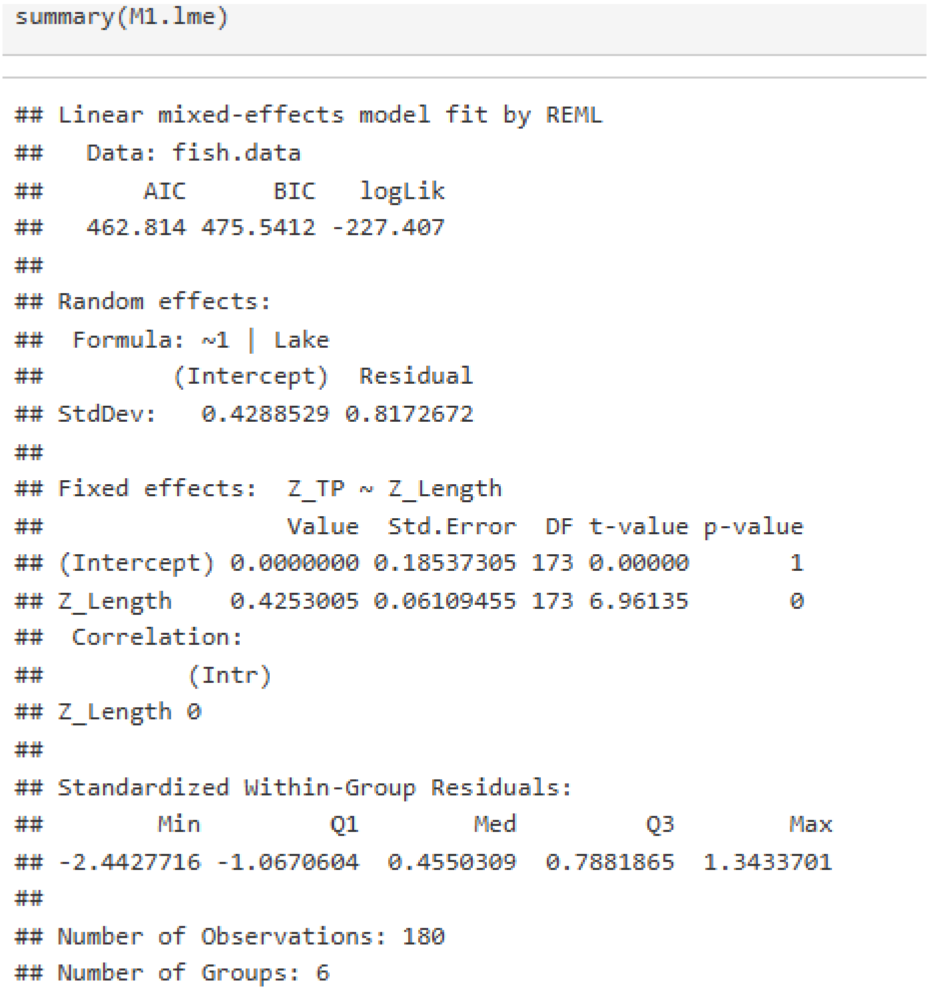
Step 9: Refit with REML and Validate



```
M1.lme= lme(Z_TP ~ Z_Length, random=~1|Lake,
     data = fish.data, method="REML")
```

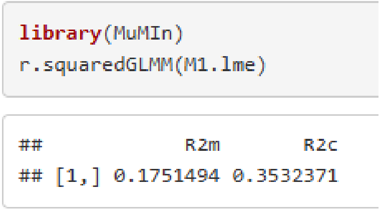
- Validate = Check assumptions
  - Our model has not changed – OK
  - If model changed, recheck assumptions

Step 10: Interpret & Plot



```
summary(M1.lme)
```

```
## Linear mixed-effects model fit by REML
##   Data: fish.data
##        AIC      BIC   logLik
##   462.814 475.5412 -227.407
##
## Random effects:
##  Formula: ~1 | Lake
##         (Intercept)  Residual
## StdDev:   0.4288529 0.8172672
##
## Fixed effects:  Z_TP ~ Z_Length
##                 Value  Std.Error  DF t-value p-value
## (Intercept) 0.0000000 0.18537305 173 0.00000       1
## Z_Length    0.4253005 0.06109455 173 6.96135       0
##  Correlation:
##          (Intr)
## Z_Length 0
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.4427716 -1.0670604  0.4550309  0.7881865  1.3433701
##
## Number of Observations: 180
## Number of Groups: 6
```



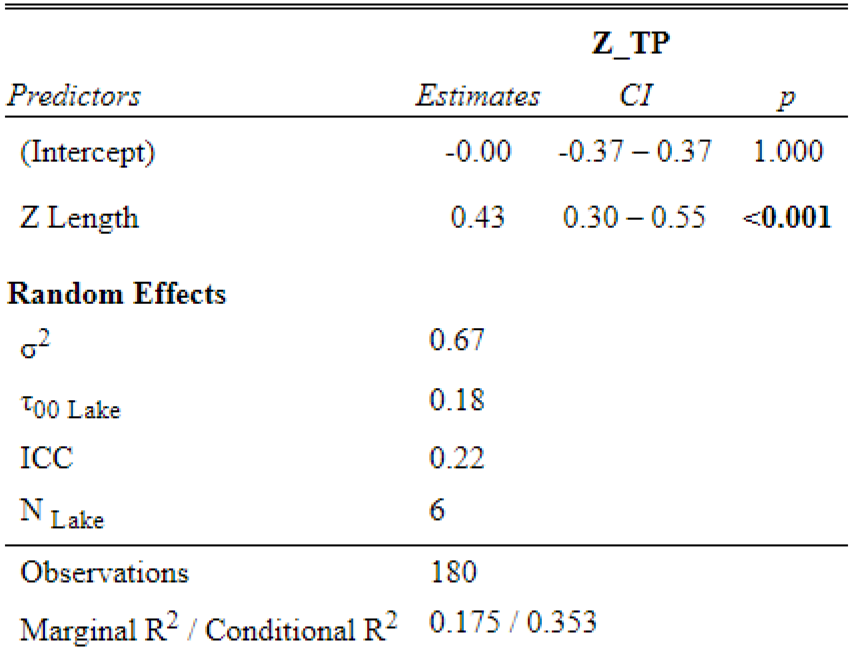
```
library(MuMIn)
r.squaredGLMM(M1.lme)
```

```
##            R2m       R2c
## [1,] 0.1751494 0.3532371
```

Random effects:

- StdDev – how much the intercepts differ across lakes

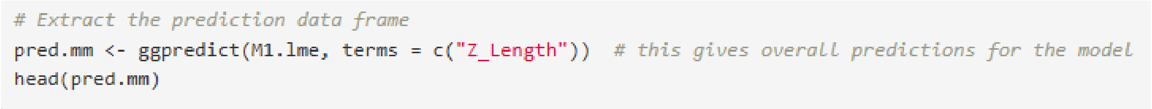
- ICC = $(0.43^2)/(0.43^2 + 0.82^2) = 0.21$
- ICC = $(0.18)/(0.67+0.18) = 0.21$



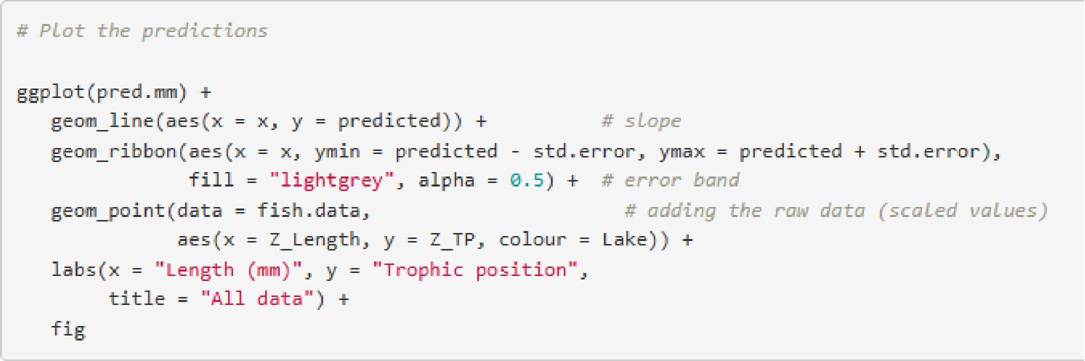
	Z_TP		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.00	-0.37 – 0.37	1.000
Z Length	0.43	0.30 – 0.55	**<0.001**
Random Effects			
σ^2	0.67		
τ_{00} Lake	0.18		
ICC	0.22		
N Lake	6		
Observations	180		
Marginal R^2 / Conditional R^2	0.175 / 0.353		

R^2 values show how much variance is explained by fixed (marginal) and fixed+random (Conditional)

Step 10: Interpret & Plot

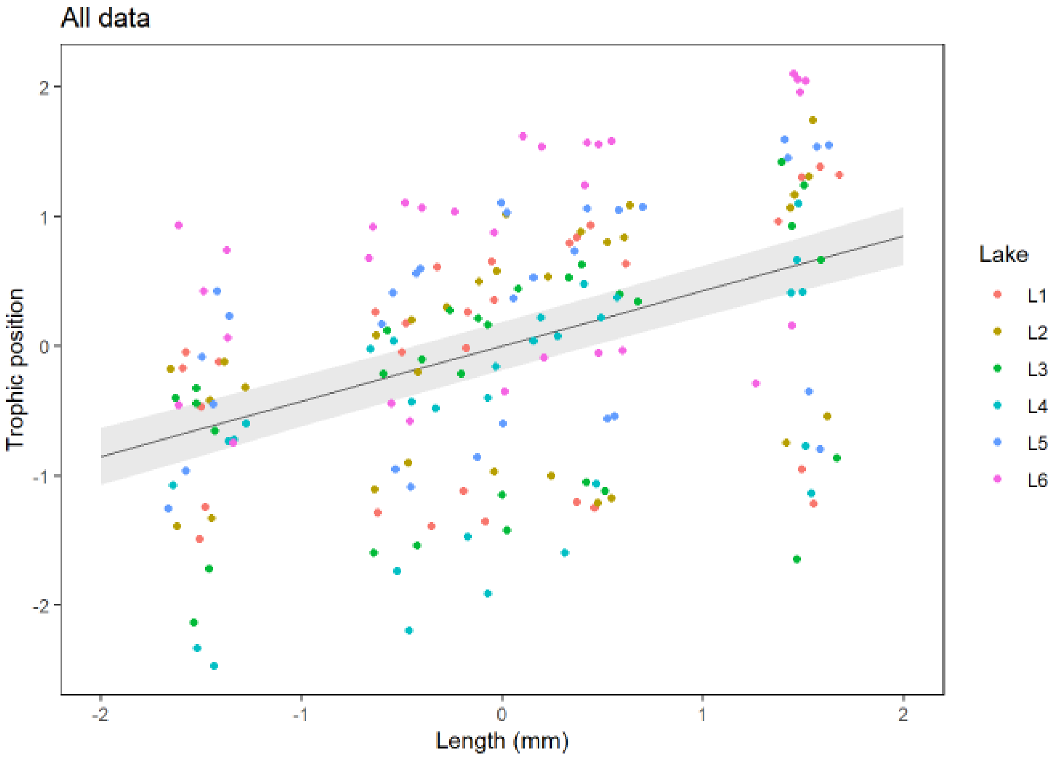


```r
# Extract the prediction data frame
pred.mm <- ggpredict(M1.lme, terms = c("Z_Length"))  # this gives overall predictions for the model
head(pred.mm)
```



```r
# Plot the predictions

ggplot(pred.mm) +
   geom_line(aes(x = x, y = predicted)) +          # slope
   geom_ribbon(aes(x = x, ymin = predicted - std.error, ymax = predicted + std.error),
               fill = "lightgrey", alpha = 0.5) +  # error band
   geom_point(data = fish.data,                      # adding the raw data (scaled values)
              aes(x = Z_Length, y = Z_TP, colour = Lake)) +
   labs(x = "Length (mm)", y = "Trophic position",
        title = "All data") +
   fig
```



LMM: review

1. Fixed effects: groups of interest (e.g., the overall effect of a treatment)
2. Random effects: groups are a subset of realized possibilities (e.g., experimental subjects, study sites, or time points).
3. LMM account for non-independence of observations (nested, crossed)
4. The impact of a random effect can be measured with ICC
5. Additional model assumptions: random intercepts and slopes must be normally distributed, and groups need to be independent.
6. Report both *marginal* and *conditional* R^2