Intro to mixed effects models

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

4 May 2020

Goals for today

- Understand types of random effects structures
- Understand how random effects are estimated
- Understand restricted maximum likelihood
- Understand approaches to make inference from mixed models

Model for means

Imagine we are interested in modeling the mass of fish measured in several different lakes

We have 3 hypotheses about the variation in fish sizes

1. differences in mass are due mostly to individual fish with no differences among lakes

Model for means

Imagine we are interested in modeling the mass of fish measured in several different lakes

We have 3 hypotheses about the variation in fish sizes

1. differences in mass are due mostly to individual fish with no differences among lakes

2. differences in mass are due mostly to *specific* factors that differ among lakes

Model for means

Imagine we are interested in modeling the mass of fish measured in several different lakes

We have 3 hypotheses about the variation in fish sizes

1. differences in mass are due mostly to individual fish with no differences among lakes
2. differences in mass are due mostly to *specific* factors that differ among lakes
3. differences in mass are due mostly to *general* factors that are shared among lakes

Model for means

Our first model simply treats all of the fish i in the different lakes j as one large group

$$y_{ij} = \mu + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \text{N}(0, \sigma_\epsilon^2)$$

where μ is the mean mass of fish across *all* lakes & our primary interest is the size of σ_ϵ^2

Model for means

In essence, we are *pooling* all of fish from the different lakes together so we can drop the j subscript

$$
\begin{gathered}
y_{ij} = \mu + \epsilon_{ij} \\
\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2) \\
\Downarrow \\
y_i = \mu + \epsilon_i \\
\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)
\end{gathered}
$$

Model for means

Our second model separates all of the fish i into groups based on the *specific* lake j from which they were caught

$$
y_{ij} = \mu + \alpha_j + \epsilon_{ij}
$$
$$
\epsilon_{ij} \sim \text{N}(0, \sigma_\epsilon^2)
$$

where α_j is the *specific* effect of lake j

Model for means

Here there is *no pooling* of fish from different lakes and the j subscript tells us about a *specific* lake

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2)$$

Model for means

Our last model treats differences in fish mass among lakes as similar to one another (correlated)

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_j + \epsilon_{ij} \\
\epsilon_{ij} &\sim \text{N}(0, \sigma^2_\epsilon) \\
\alpha_j &\sim \text{N}(0, \sigma^2_\alpha)
\end{aligned}
$$

where α_j is the effect of lake j as though it were *randomly* chosen

Model for means

The degree of correlation among lakes (ρ) is determined by the relative sizes of σ_α^2 and σ_ϵ^2

$$
\begin{gathered}
y_{ij} = \mu + \alpha_j + \epsilon_{ij} \\
\epsilon_{ij} \sim \text{N}(0, \sigma_\epsilon^2) \\
\alpha_j \sim \text{N}(0, \sigma_\alpha^2) \\
\Downarrow \\
\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}
\end{gathered}
$$

Model for means

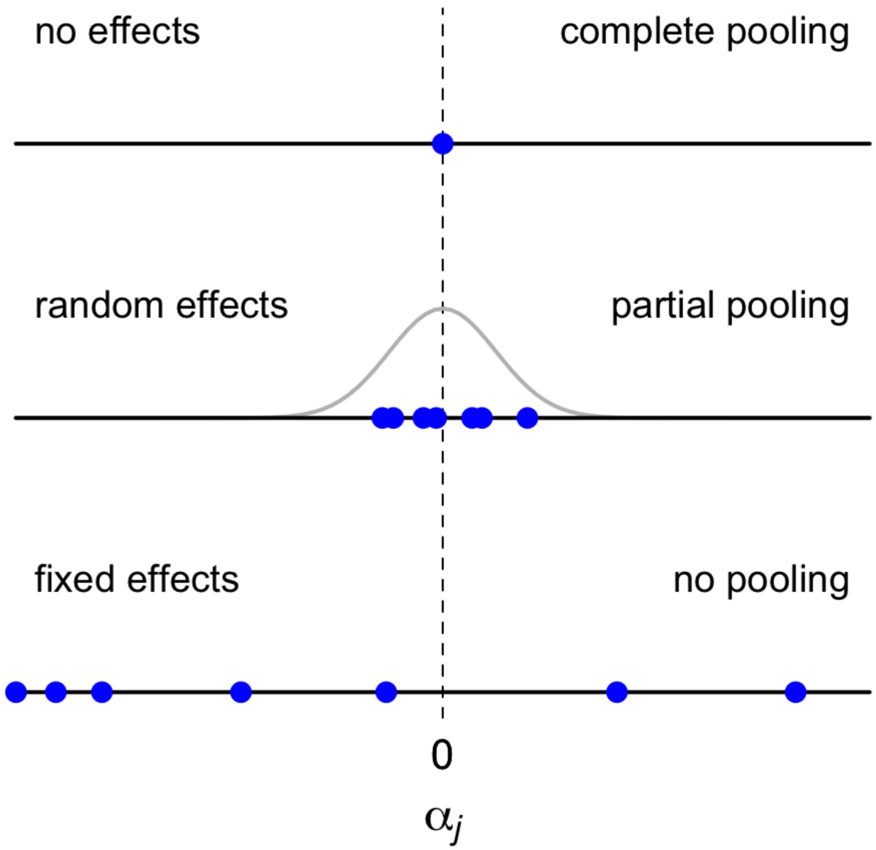
Here we could say that the lakes are *partially pooled* together by formally addressing correlations among lakes

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_j + \epsilon_{ij} \\
\epsilon_{ij} &\sim \text{N}(0, \sigma_\epsilon^2) \\
\alpha_j &\sim \text{N}(0, \sigma_\alpha^2)
\end{aligned}
$$

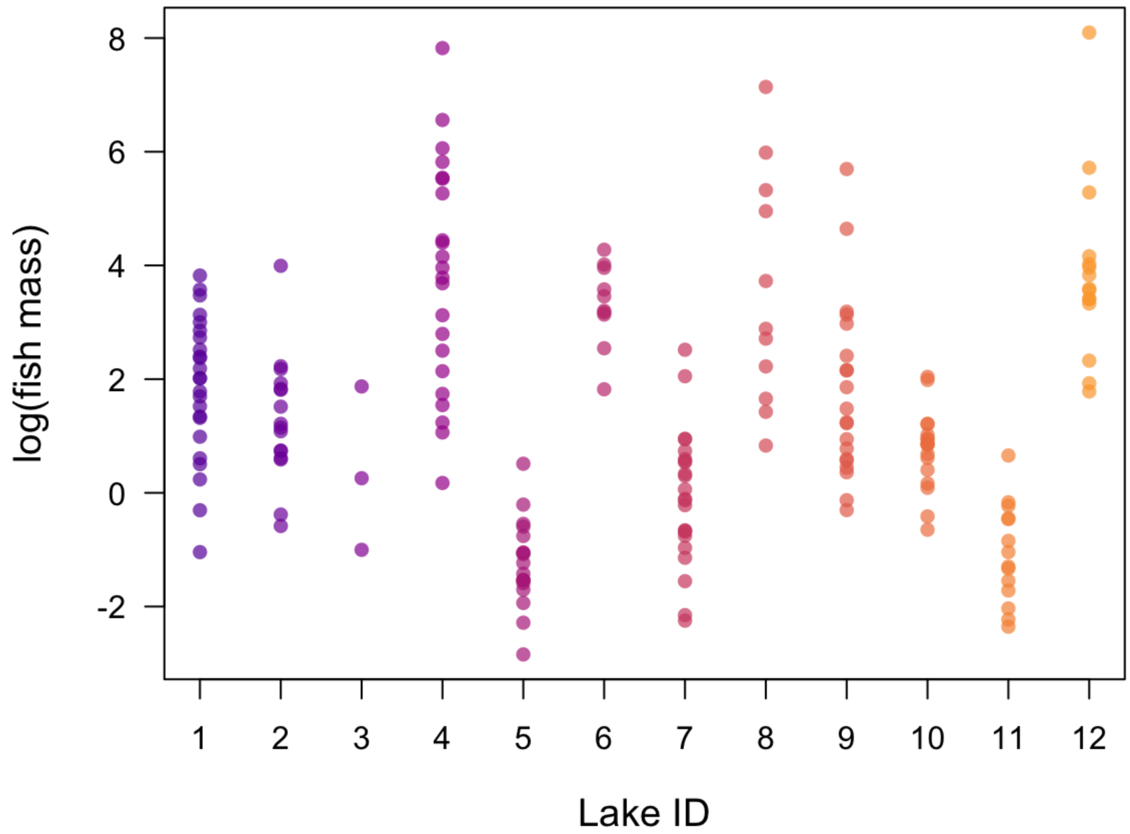
with

$$
\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}
$$

Model for means



Fish mass across lakes

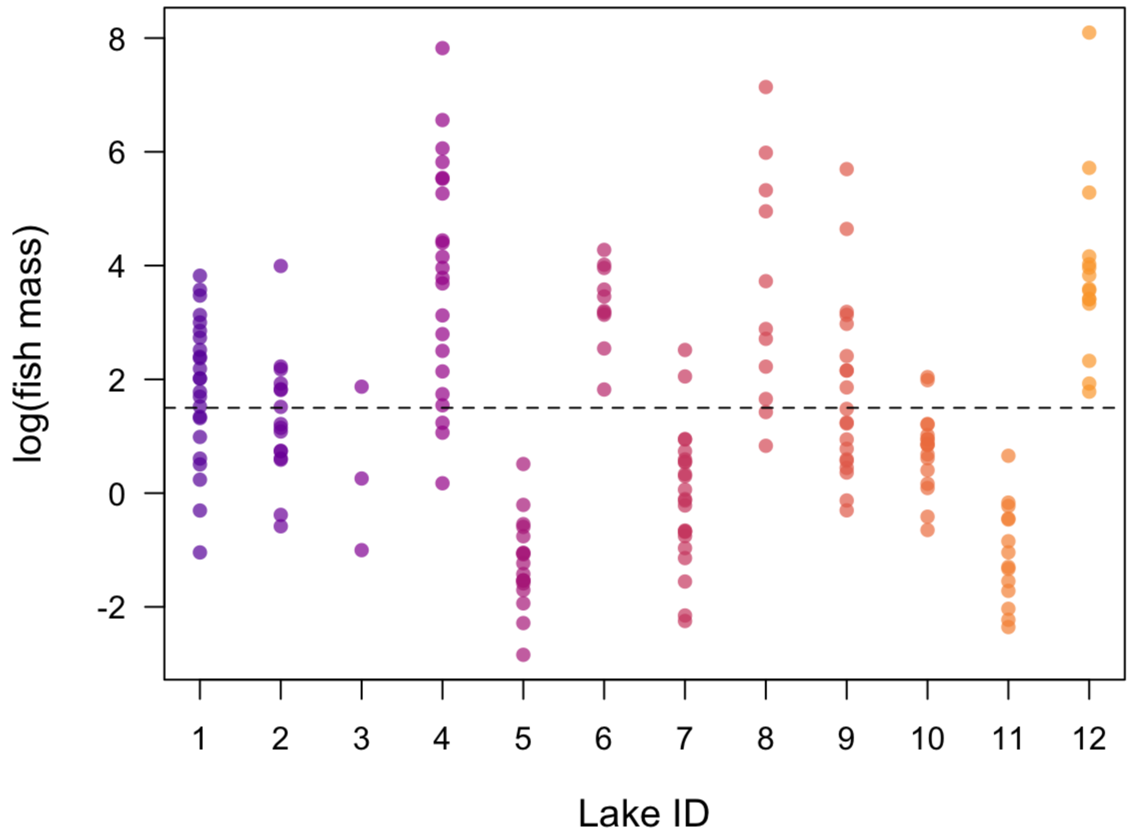


Fish mass across lakes

Simple model with complete pooling

```
## log of fish mass (lfm) as grand mean
m1 <- lm(lfm ~ 1)
```

Fish mass across lakes

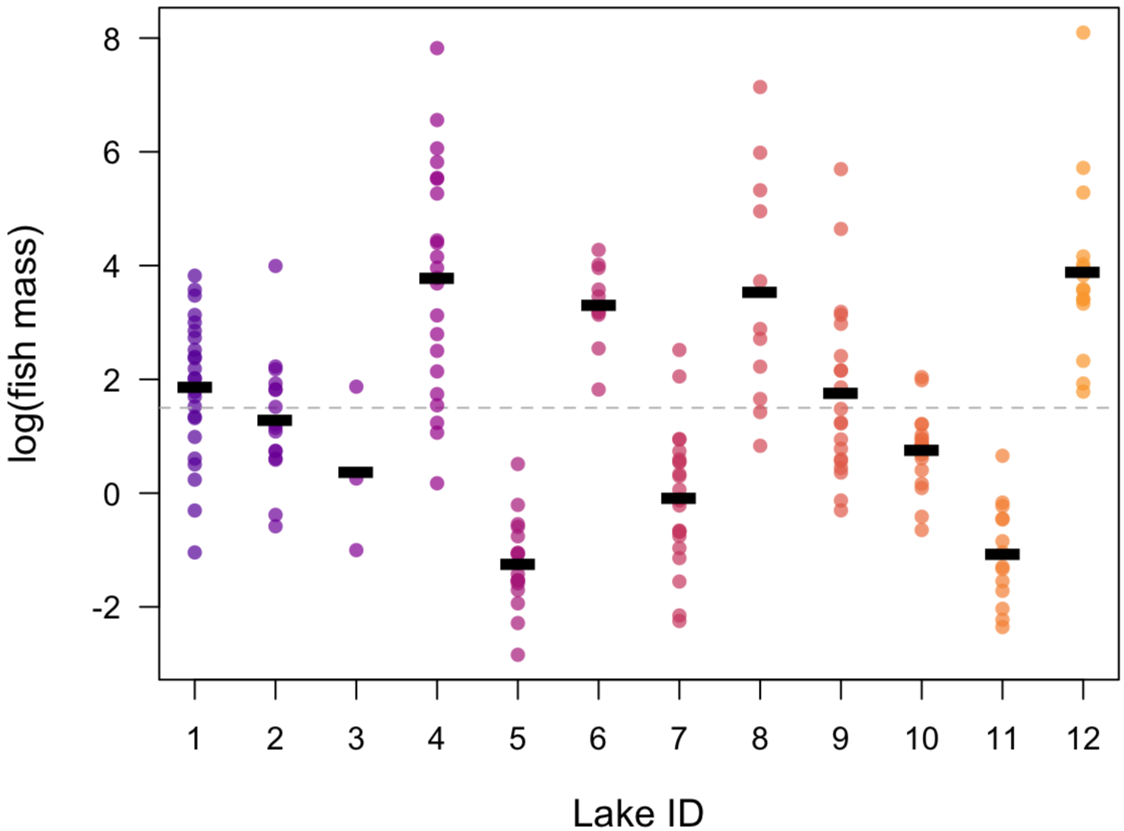


Fish mass across lakes

Fixed effects model with no pooling across lakes

```
## log of fish mass (lfm) with lake-level means
m2 <- lm(lfm ~ 1 + as.factor(IDs))
```

Fish mass across lakes

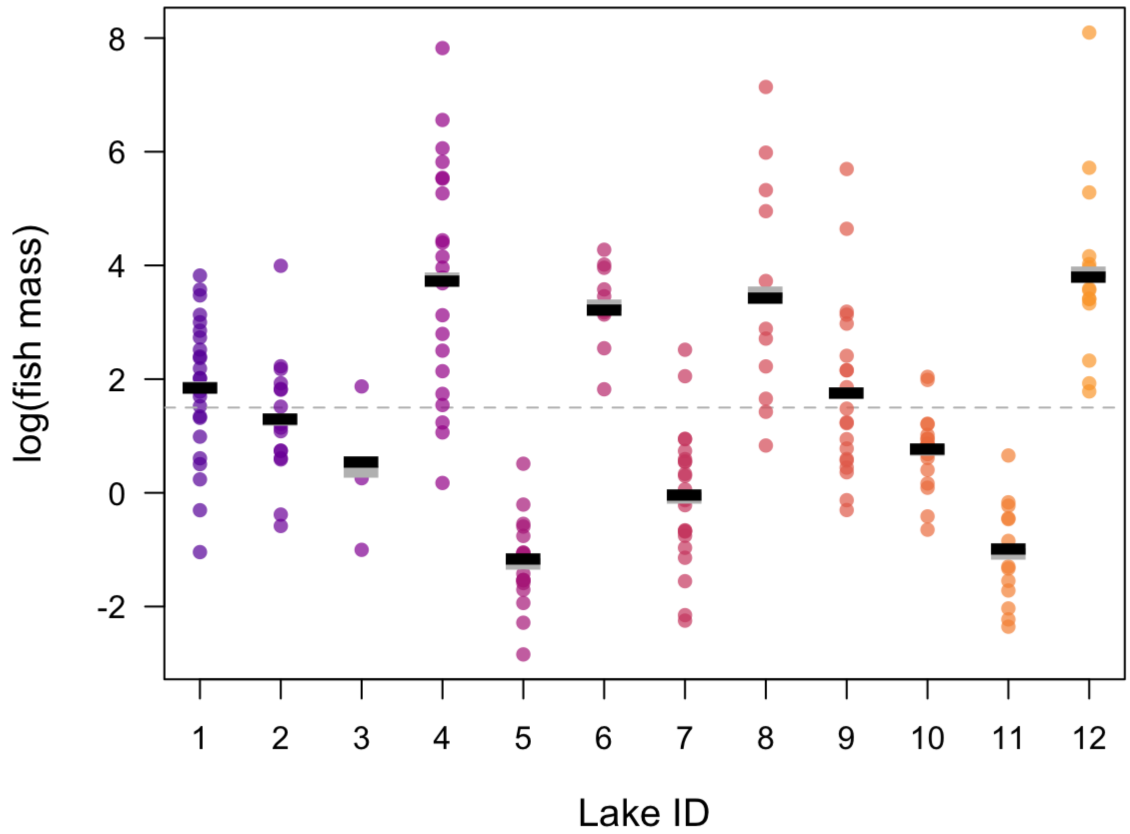


Fish mass across lakes

Random effects model with partial pooling across lakes

```
## load lme4 package
library(lme4)
## log of fish mass (lfm) with lake-level effects
m3 <- lmer(lfm ~ 1 + (1|IDs))
```

Fish mass across lakes



Shrinkage of group means

In fixed effects models, the group means are

$$\alpha_j = \bar{y} - \mu$$

In random effects models, the group means “shrink” towards the mean

$$\alpha_j = (\bar{y} - \mu)\left(\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}\right)$$

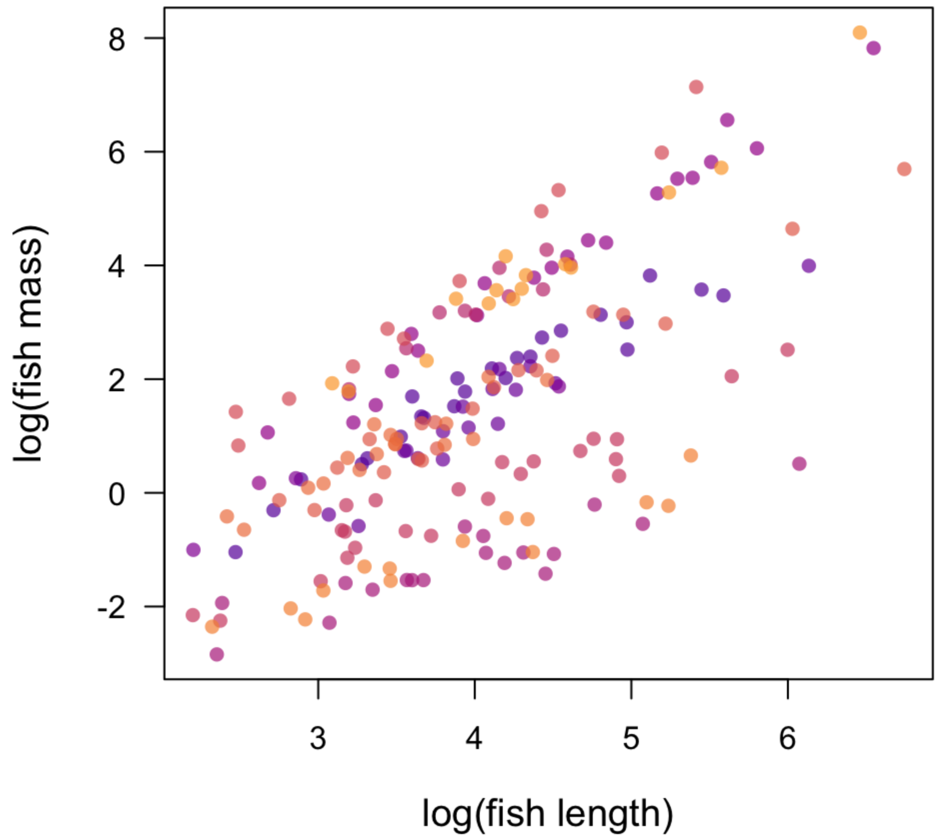
QUESTIONS?

Fish mass across lakes

Let’s return to our model for fish mass across different lakes

Now we want to include the effect of fish length as well

Fish mass versus length



A global regression model

Fish mass as a function of its length (no lake effects)

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{fixed}} + \epsilon_{ij}$$

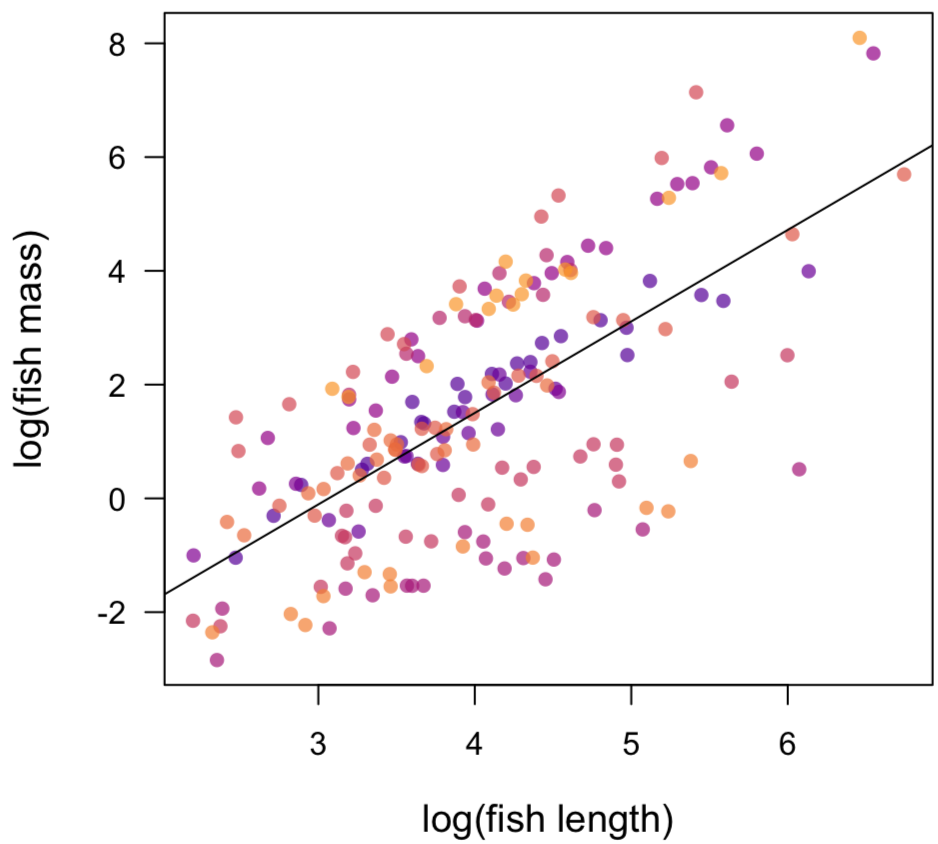
$\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon)$

A global regression model

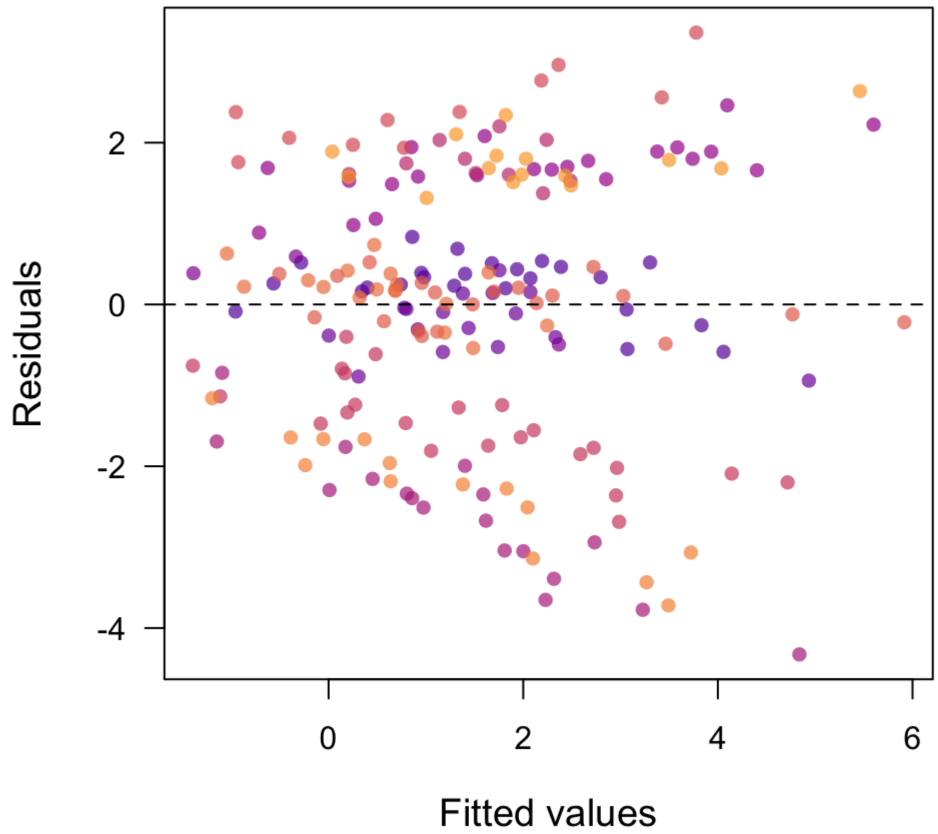
Fish mass as a function of its length (no lake effects)

```
## fit global regression model
a1 <- lm (lfm ~ lfl)
```

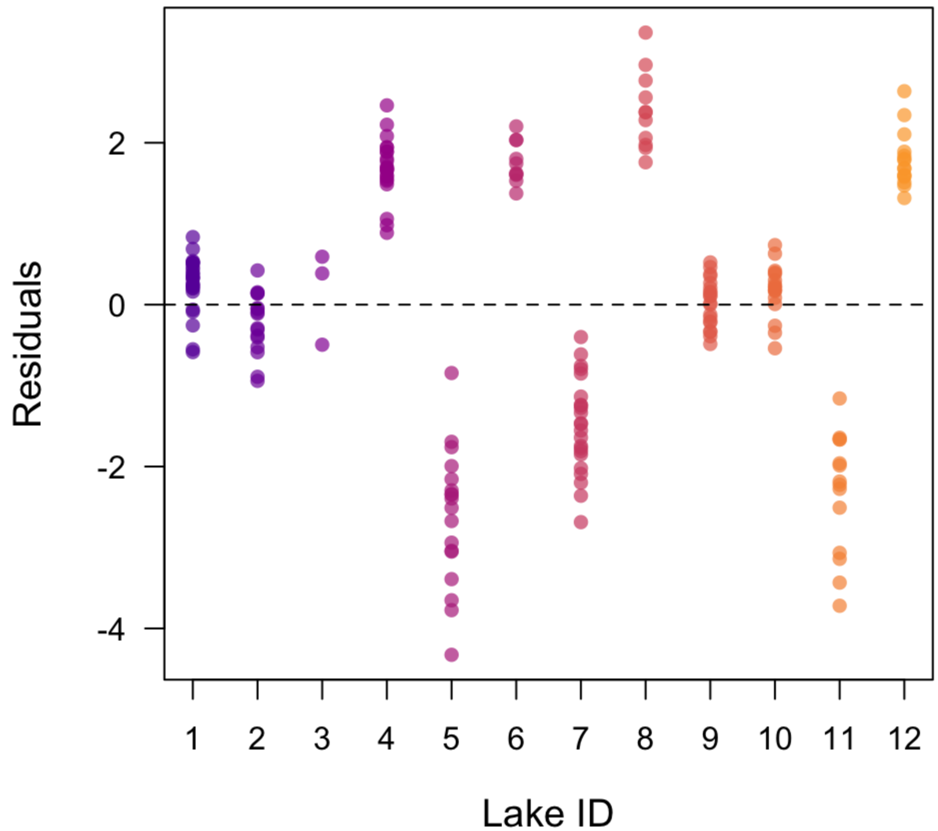
A global regression model



A global regression model



A global regression model



Unique regression models

Fish mass as a function of its length for *each* lake

$$y_{ij} = \underbrace{\beta_{0j} + \beta_{1j} x_{ij}}_{\text{fixed}} + \epsilon_{ij}$$

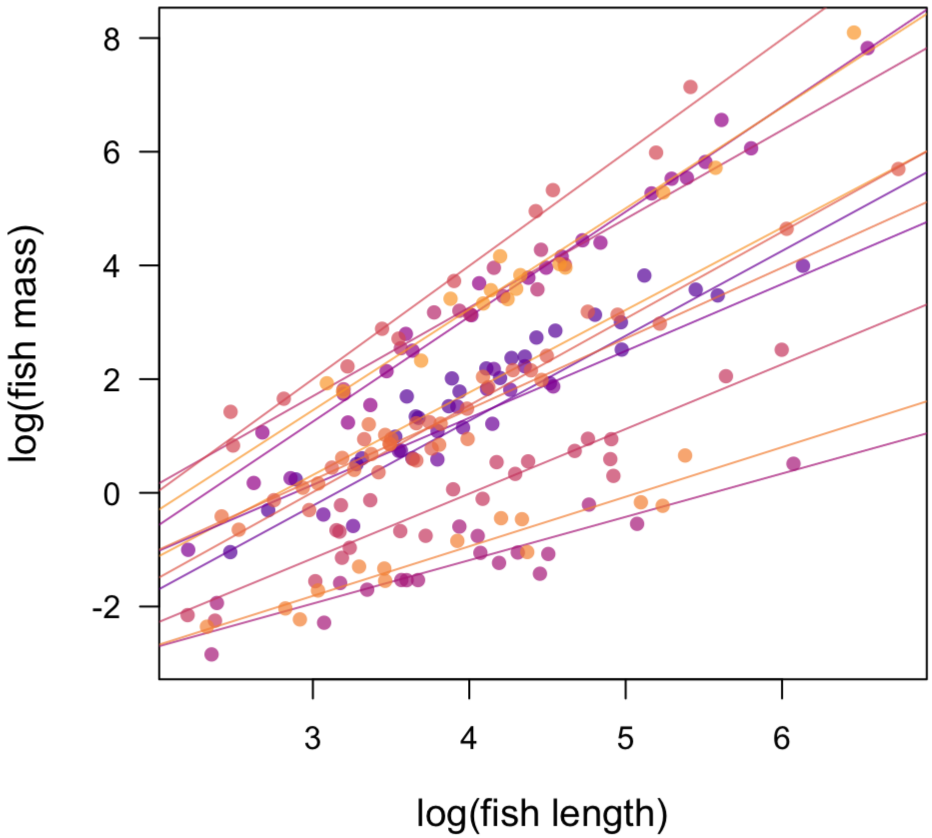
$\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon)$

Unique regression models

Fish mass as a function of its length for *each* lake

```
## matrix for coefs
cf <- matrix(NA, nl, 2)
## fit regression unique to each lake
for(i in 1:nl) {
  cf[i,] <- coef(lm(fm[[i]] ~ fl[[i]]))
}
```

Unique regression models



A linear mixed model

Fish mass as a function of its length for a *random* lake

$$y_{ij} = \underbrace{\beta_{0j} + \beta_1 x_{ij}}_{\text{fixed}} + \underbrace{\alpha_j}_{\text{random}} + \epsilon_{ij}$$

$\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon)$

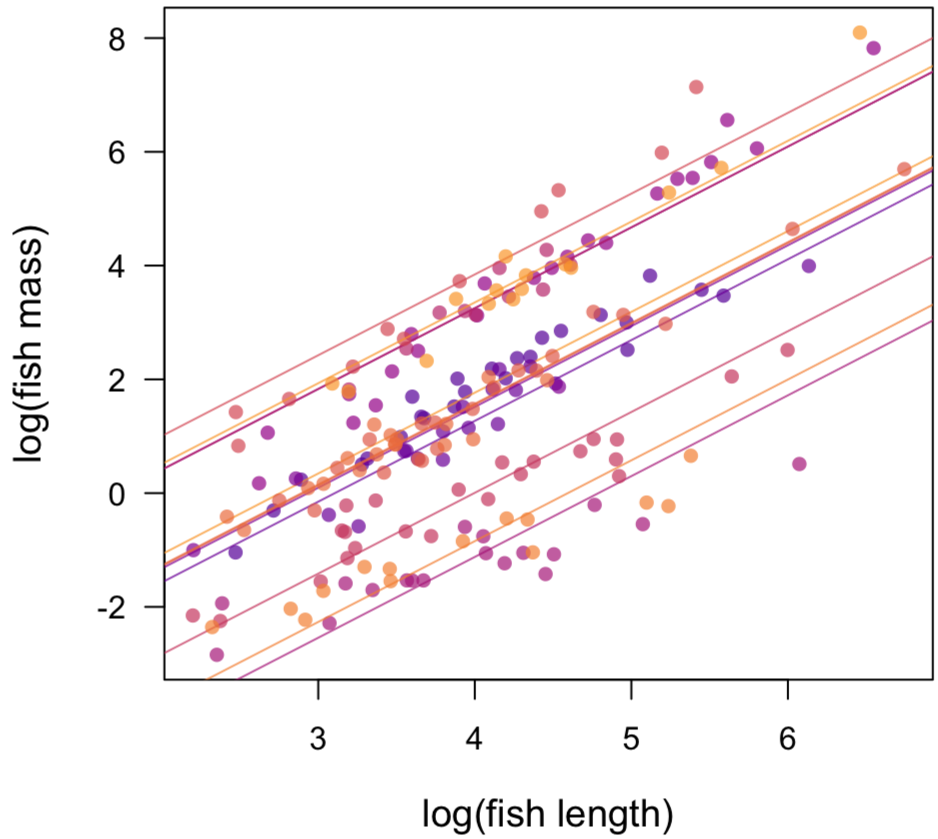
$\alpha_j \sim \mathrm{N}(0, \sigma_\alpha)$

A linear model (ANCOVA)

Fish mass as a function of its length and *random* lake

```
## fit ANCOVA with fixed factor for length & rdm factor for lake
a2 <- lmer(lfm ~ lfl + (1|IDs))
```

Fish mass versus length



A random effects model

Fish mass as a function of its length for a *random* fish *and* lake

$$y_{ij} = (\beta_{0j} + \alpha_j) + (\beta_{1j} + \delta_j)x_{ij} + \epsilon_{ij}$$
$$y_{ij} = \underbrace{\beta_{0j} + \beta_{1j}x_{ij}}_{\text{fixed}} + \underbrace{\alpha_j + \delta_j x_{ij}}_{\text{random}} + \epsilon_{ij}$$

$\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon)$

$\alpha_j \sim \mathrm{N}(0, \sigma_\alpha)$

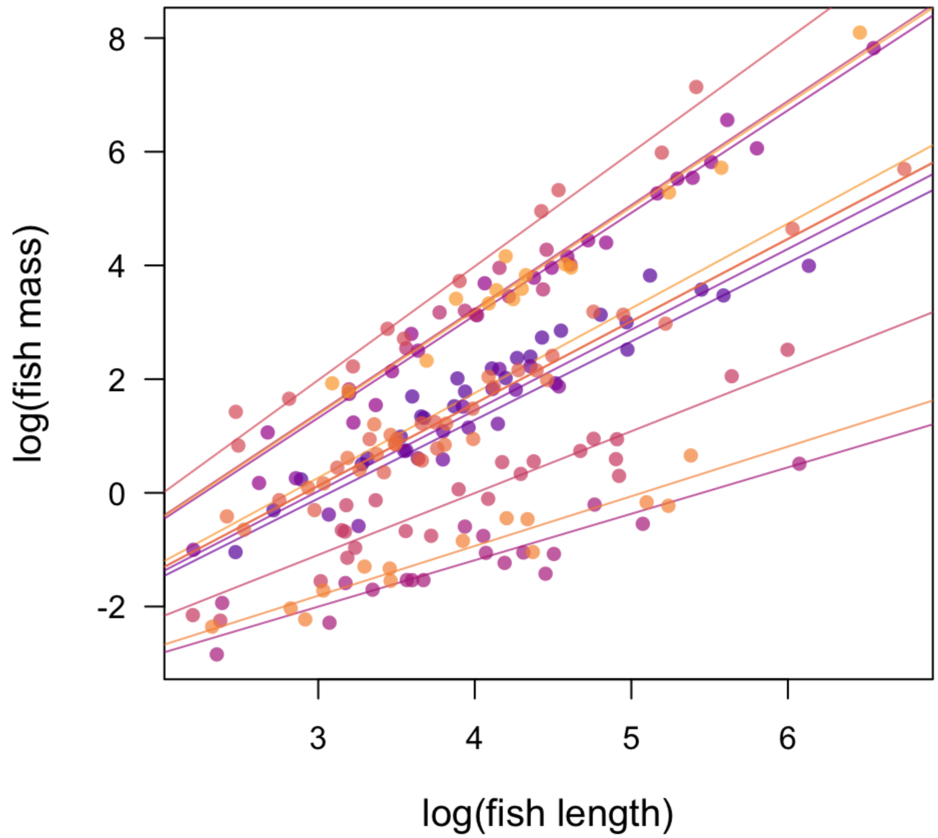
$\delta_j \sim \mathrm{N}(0, \sigma_\delta)$

A random effects model

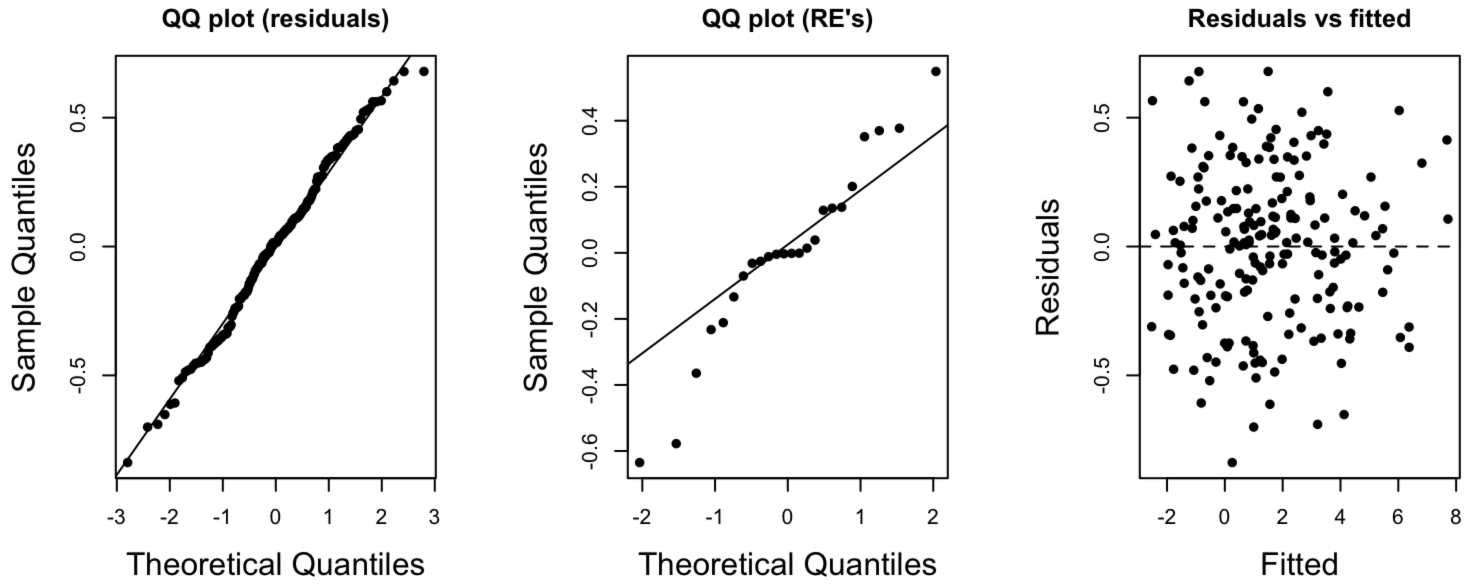
Fish mass as a function of its length for a *random* fish *and* lake

```
## fit ANCOVA with random effects for length & lake
a3 <- lmer(lfm ~ lfl + (lfl|IDs))
```

A random effects model



Model diagnostics



QUESTIONS?

General linear model

We have seen how to write a general linear model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ contains the *fixed effects* of $\mathbf{X}$ on $\mathbf{y}$

General linear mixed model

We can extend the general linear model to include both of fixed and random effects (a *mixed effects model*)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}$$

where $\mathbf{Z}$ is also a design matrix and $\mathbf{Z}$ contains a mix of $z \in \{-1, 0, 1\}$ and $z \in \mathbb{R}$

General linear mixed model

We can extend the general linear model to include both of fixed and random effects (a *mixed effects model*)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}$$

$$\mathbf{e} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$$

$$\boldsymbol{\alpha} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{D})$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{D}$ is a square matrix of constants

General linear mixed model

Variance decomposition

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}$$
$$\Downarrow$$
$$\mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) + \mathrm{Var}(\mathbf{Z}\boldsymbol{\alpha}) + \mathrm{Var}(\mathbf{e})$$

General linear mixed model

Variance of random components

$$\mathrm{Var}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}) = \mathrm{Var}(\mathbf{Z}\boldsymbol{\alpha}) + \mathrm{Var}(\mathbf{e})$$

$$\Downarrow$$

$$\begin{aligned}
\mathbf{V} &= \mathbf{Z}\mathrm{Var}(\boldsymbol{\alpha})\mathbf{Z}^\top + \mathrm{Var}(\mathbf{e}) \\
&= \mathbf{Z}(\sigma^2\mathbf{D})\mathbf{Z}^\top + \sigma^2\mathbf{I} \\
&= \sigma^2(\mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \mathbf{I})
\end{aligned}$$

Log-likelihood for fixed effects

Recall that we think of likelihoods in terms of the *observed data*

But the random effects in our model are *unobserved* random variables, so we need to integrate them out of the likelihood

Log-likelihood for fixed effects

The log-likelihood for the fixed effects $\boldsymbol{\beta}$

$$\log \mathcal{L}(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Estimate of fixed effects

This leads us to our familiar statement for the weighted least squares estimate for $\boldsymbol{\beta}$

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \min \; (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}) \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y}
\end{aligned}
$$

Variance of fixed effects

Our variance estimate for $\boldsymbol{\beta}$ is then

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}$$

Log-likelihood for random effects

The log-likelihood for the random effects is given by

$$\begin{aligned}\log \mathcal{L}(\mathrm{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{\sigma^2}{2} - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}) \\ - \frac{1}{2}\left|\mathbf{Z}\mathbf{D}\mathbf{Z}^\top\right| - \frac{1}{2}\boldsymbol{\alpha}^\top(\mathbf{Z}\mathbf{D}\mathbf{Z}^\top)^{-1}\boldsymbol{\alpha}\end{aligned}$$

Estimate of random effects

This leads to the *best linear unbiased predictor* for $\boldsymbol{\alpha}$

$$\hat{\boldsymbol{\alpha}} = \sigma^2(\mathbf{Z}\mathbf{D}\mathbf{Z}^\top)\mathbf{Z}^\top\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Restricted maximum likelihood

Estimating the parameters in a mixed effects model requires *restricted maximum likelihood* (REML)

REML works by

1. estimating the fixed effects $(\hat{\boldsymbol{\beta}})$ via ML

2. using the $\hat{\boldsymbol{\beta}}$ to estimate the $\hat{\boldsymbol{\alpha}}$

lme4 makes this easy for us

Inference for mixed models

With random effects models, we can't use our standard inference tools because we don't know the distributions for our test statistic

(**lme4** won't give p-values)

Inference for mixed models

Likelihood ratio test

We can use a likelihood ratio test for nested models, but the assumption of a χ^2 distribution can be poor

Inference for mixed models

F test

We can also use F-tests to evaluate a single fixed effect, but again the assumption of a F distribution can be poor

Inference for mixed models

Bootstrapping

We can use bootstrapping to conduct likelihood ratio tests

1. simulate data from the simple model

2. fit simple & full model and calculate likelihood ratio

3. see where test statistic falls within estimated distribution from (2)

Inference for mixed models

We can report parameter estimates and CI’s via bootstrapping

We can generate predictions given fixed and random effects and estimate their uncertainty via bootstrapping

Model selection

Recall that $AIC = 2k - 2\log\mathcal{L}$

The problem with mixed effects models is that it's not clear what k equals

It works well to select among fixed effects if random effects are held constant

Model selection

To use AIC, we can follow these steps

1. Fit a model with *all* of the possible fixed-effects included
2. Keep the fixed effects constant and search for random effects
3. Keep random effects as is and fit different fixed effects

Model selection

Other options include

- BIC
- cross-validation

Summary

- Think hard about your question and data
  - are there groups or levels?
  - are the temporal or spatial components?
- Decide what random effects make sense
- Once random effects are chosen, select fixed effects
- Inference will generally require bootstrapping