Mixed models – aka mixed effect models, hierarchical models, multilevel models

Statistical necessity

Biological data often violates the assumption of independently distributed data

- multiple measurements on the same organism
- experiments organized into spatial blocks
- observational data where multiple observations were made at each of many locations
- observational data where multiple observations were made at each of many timepoints
- community data with multiple counts of each of many species
- data syntheses of similar experiments that were performed by many different researchers

These are all kinds of **pseudoreplication**

Mixed models can account for these kinds of structure

Mixed models – aka mixed effect models, hierarchical models, multilevel models

Biological necessity

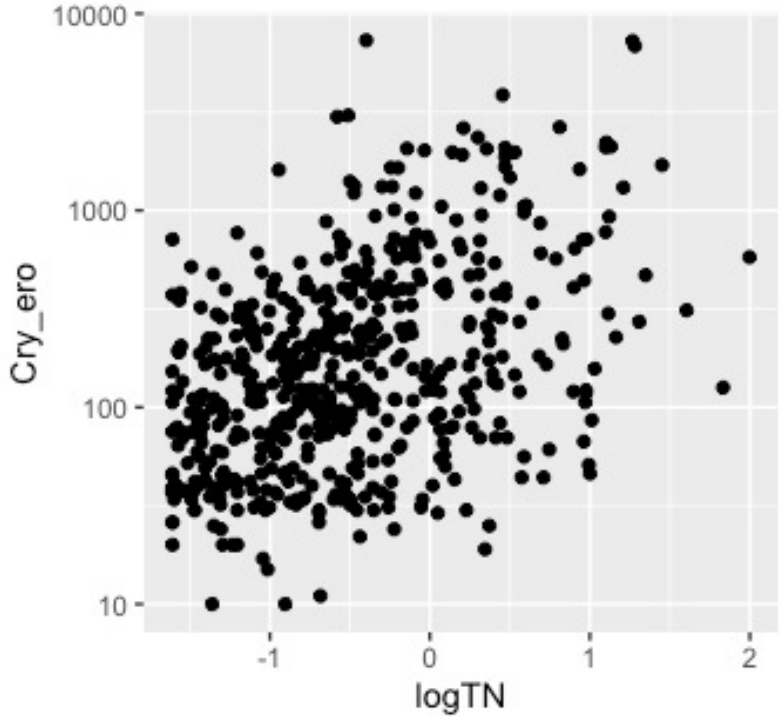
- Biological processes have multiple sources of variation, at multiple spatial + temporal scales
- We would like to model this variation, and see what explains variation at different scales

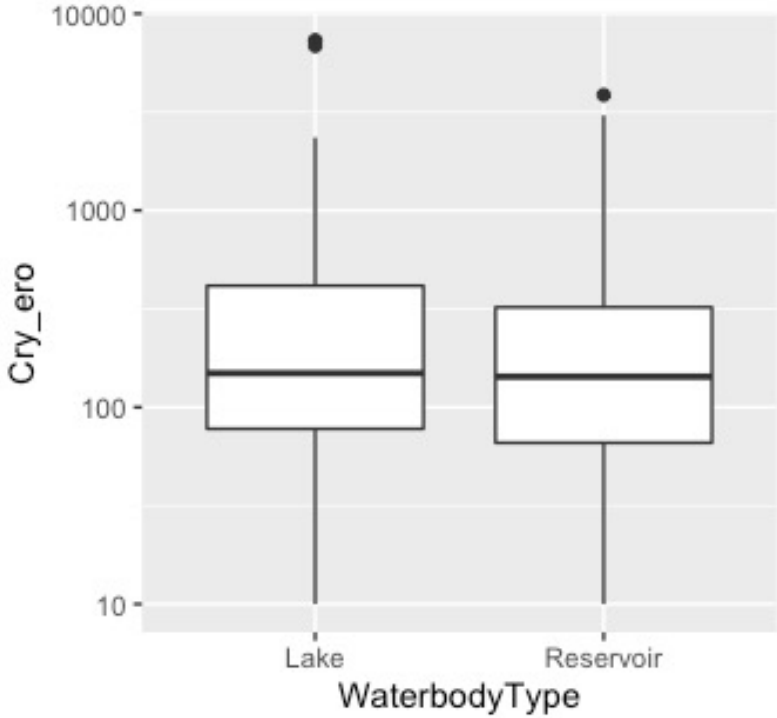
Mixed models are very good at this

An example of random effects: EPA phytoplankton samples from many lakes

- Community samples in >500 lakes, 1-4 measurements per lake (spring, summer, fall)
- Let's model the abundance of *Cryptomonas erosa*

- Raw data:

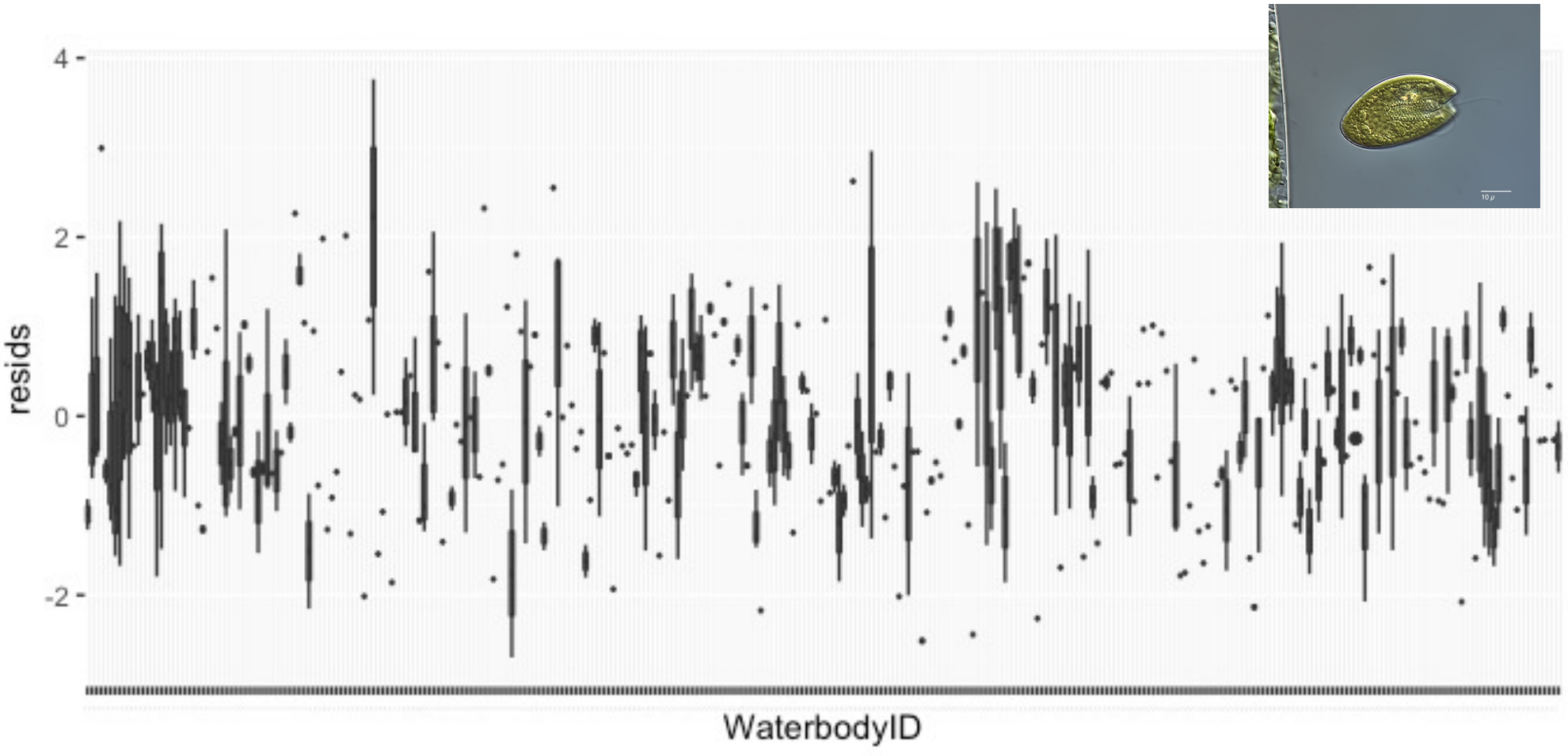




- If we just use a linear model, we ignore the spatial structure
- Can average within lakes, but lose TN variation, seasonal signal

An example of random effects: EPA phytoplankton samples from many lakes

```
crysub$resids = resid(lm(log(Cry_ero) ~ logTN, data = crysub))
ggplot(crysub, aes(WaterbodyID, resids)) + geom_boxplot() + theme(axis.
text.x = element_text(angle = 45, vjust = -1, hjust=1, size = 0.5))
```



- The data are not independently distributed

How can we account for variation among 322 waterbodies?

We could add a factor for WaterbodyID to the linear model

- But this would add 321 parameters (for 535 samples)

- Also we can't include a factor for WaterbodyID, as well as a predictor that varies **at the scale of waterbodies (**WaterbodyID and WaterbodyType are **collinear)**

- We would like to model variability among groups of data (waterbodies), while also asking what predicts that variation: **can't do it with LMs or GLMs**

- Mixed models allow us to have **predictors at multiple scales**

The logic of random effects

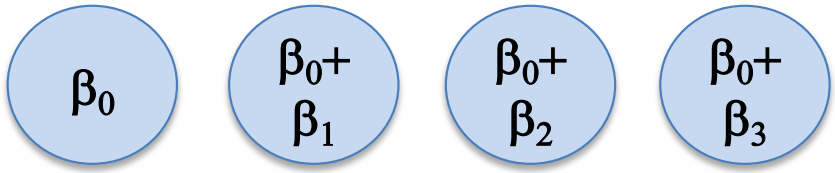
Random effects:

- Rather than fitting a separate parameter for each waterbody, we are going to assume that the variation in waterbodies is normally distributed

- And we will **estimate the mean and variance of that distribution**

Fixed effect factor

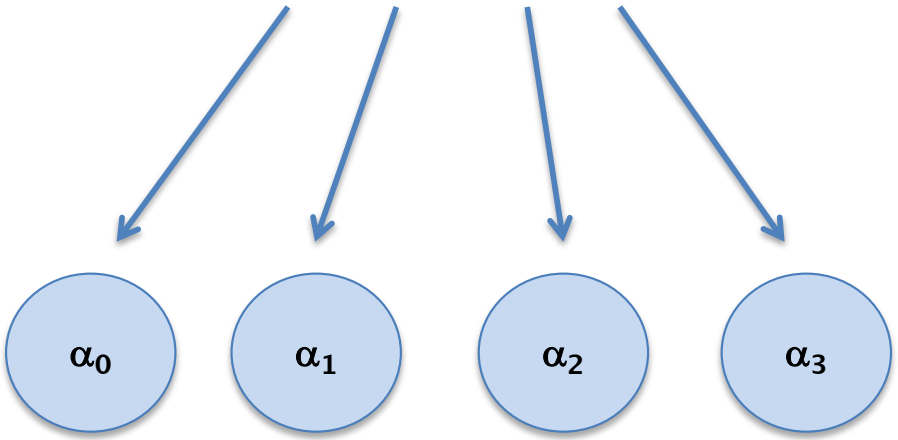
Each lake gets its own separate parameter



Random effect factor

Lakes are drawn from a common distribution:

Normal($\boldsymbol{\mu}_\alpha$, $\boldsymbol{\sigma}_\alpha$)



Different ways to write a random effects model

Fixed effects version

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$Y_i \sim \text{Normal}(\mu_i, \sigma)$$

Random effects version 1

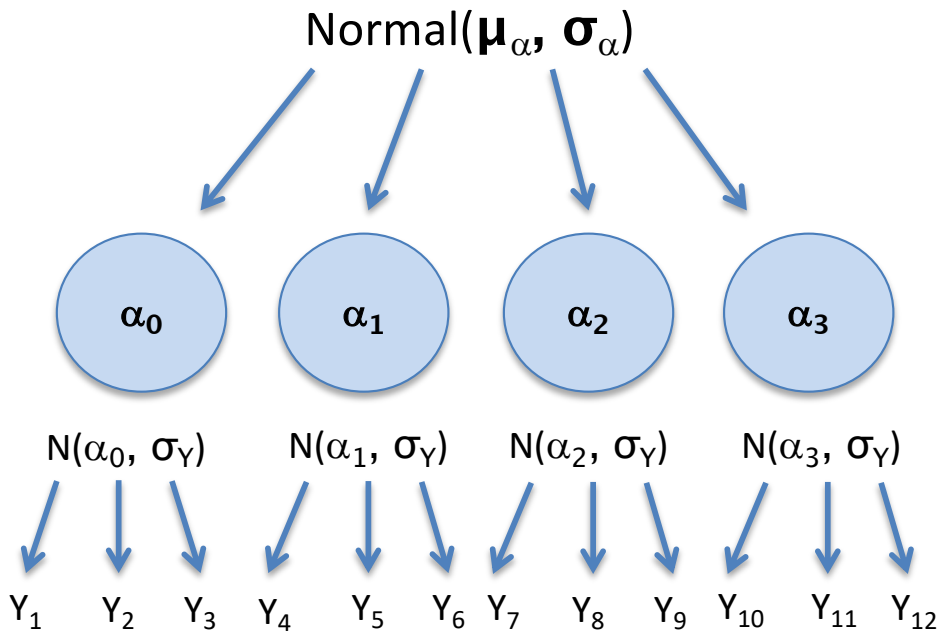
$$\mu_i = \beta_0 + b_1 Z_{1i} + b_2 Z_{2i} + b_3 Z_{3i}$$
$$b_j \sim \text{Normal}(0, \sigma_b)$$
$$Y_i \sim \text{Normal}(\mu_i, \sigma_Y)$$

Random effects version 2

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma_\alpha)$$
$$Y_i \sim \text{Normal}(\alpha_{j[i]}, \sigma_Y)$$

Random effects version 2

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma_\alpha)$$
$$Y_i \sim \text{Normal}(\alpha_{j[i]}, \sigma_Y)$$



- We are assuming a **hierarchy** of stochastic processes
- Hence the name hierarchical or multilevel models
- Biologists tend to use **'mixed effects',** because random + fixed

How to fit a varying-intercept model in R

Using a subset of 100 observations

Fixed effects version

```
fixed.mod = lm(log(Cry_ero) ~ WaterbodyID, data = crysub1)
```

Random effects version

```
library(lme4)
```

```
rand.mod = lmer(log(Cry_ero) ~ 1 + (1|WaterbodyID), data = crysub1)
```

- Remember, '1' means intercept

```
summary(rand.mod)

## Linear mixed model fit by REML ['lmerMod']
## Formula: log(Cry_ero) ~ 1 + (1 | WaterbodyID)
##    Data: crysub1
##
## REML criterion at convergence: 326.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6857 -0.7575 -0.0365  0.6161  2.1561
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  WaterbodyID (Intercept) 0.352    0.593
##  Residual                1.213    1.101
## Number of obs: 100, groups:  WaterbodyID, 48
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    5.035      0.142    35.5
```

Linear mixed model = LMM

WaterbodyID Std.Dev = $\boldsymbol{\sigma}_{\alpha}$

Residual Std.Dev = $\boldsymbol{\sigma}_{Y}$

(Intercept) = $\boldsymbol{\mu}_{\alpha}$

- variation among waterbodies is about 25% of the total variation

Random effects ‘estimates’

- Strictly speaking, this model only has 3 parameters
- But we would still like to know what the waterbody means are
- We can use the fitted model to get the ‘best’ estimates for the random effects
- Kind of magic: we get estimates without fitting a coefficient

```
ranef(rand.mod)

## $WaterbodyID
##     (Intercept)
## 401   -0.315713
## 402    0.089169
## 403   -0.004237
## 404    0.667615
## 405   -0.388181
## 406   -0.331948
## 410   -0.477245
## 411    0.079690
## 501    0.060166
## 502    0.403596
## 503   -0.138564
## 504   -0.045408
…
```

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma_\alpha)$$

$$\eta_j \sim \text{Normal}(0, \sigma_\alpha)$$

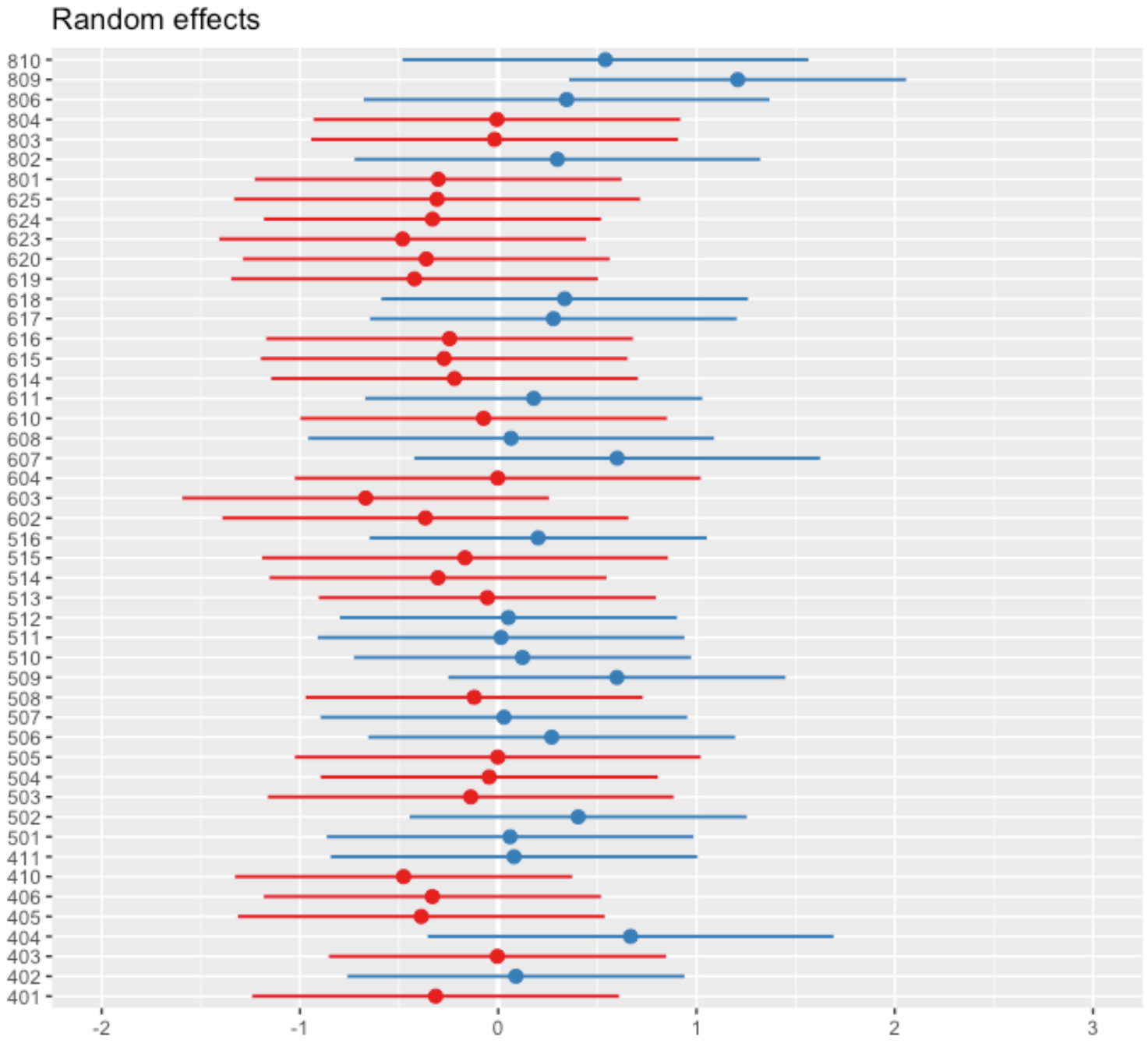
$$\alpha_j = \mu_\alpha + \eta_j$$

$\boldsymbol{\eta_j}$ = "conditional modes" of the random effects

Given our parameter estimates, and the data, what are the most likely values?

```
library(sjPlot)

plot_model(rand.mod, type = "re")
```



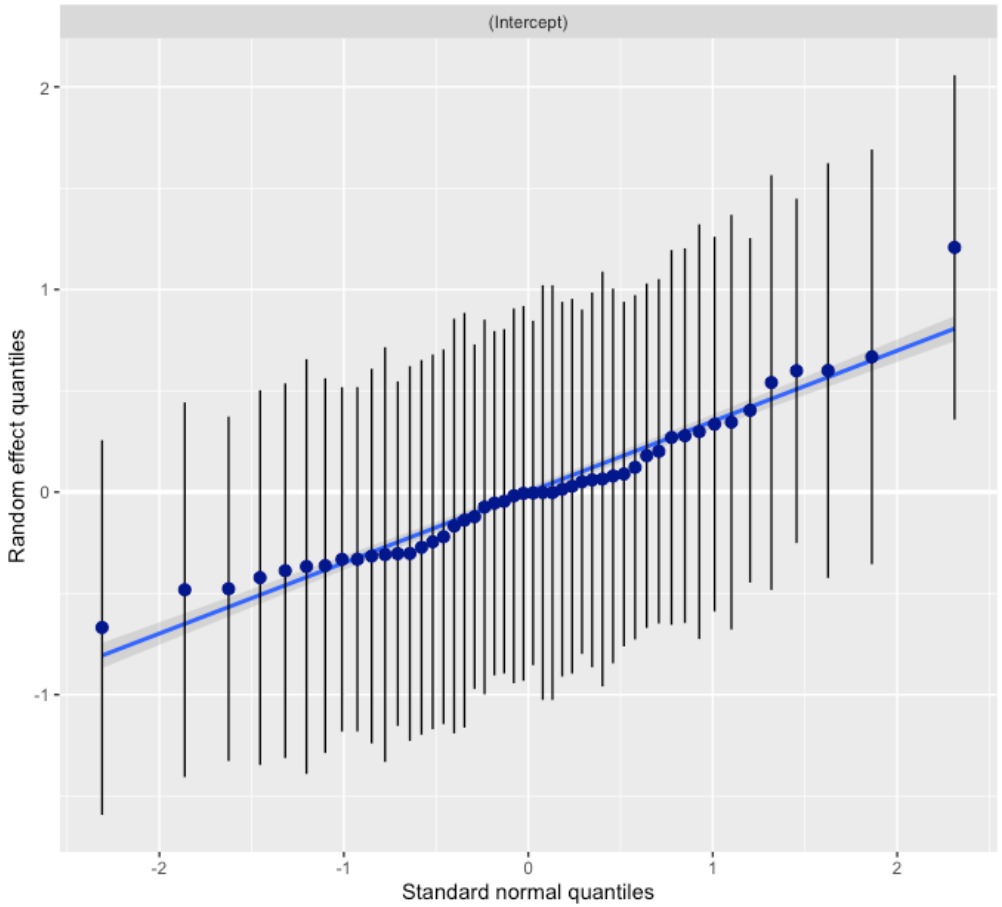
Mixed effects diagnostics

What are the assumptions for the random effects?

Can think of the random effects as residual variation at a higher level

- They should be normally distributed
- They are influenced by the model, but we can still look for outliers, etc.

```
plot_model(rand.mod, type = 'diag')[[2]]
```



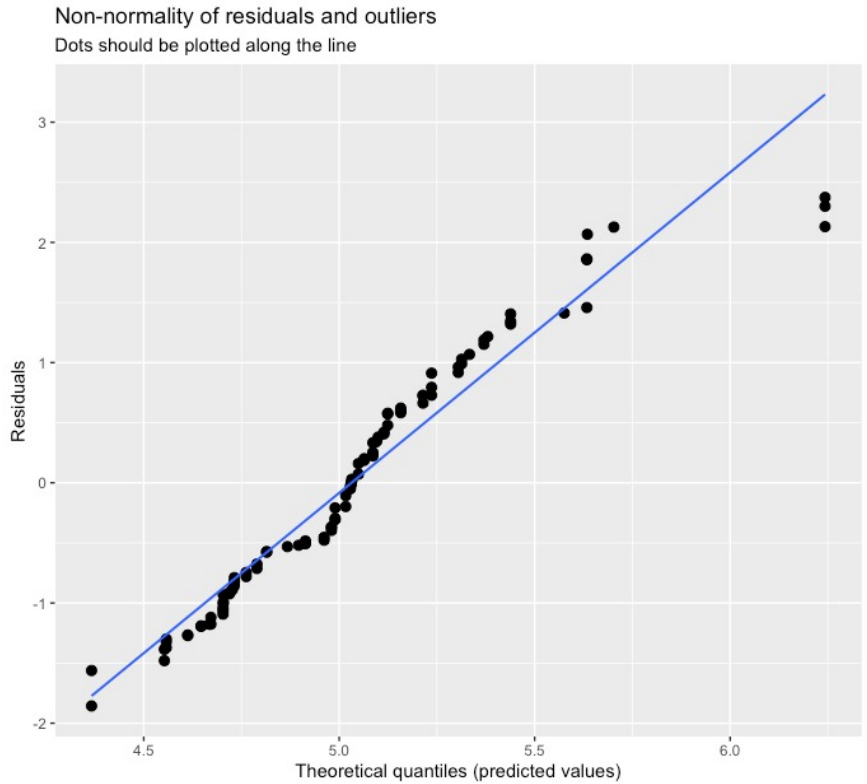
Mixed effects diagnostics

'Residuals' are the the lowest-level variation remaining after accounting for fixed and random effects

residuals() and fitted() uses the conditional modes to calculate conditional residuals

These don't have the simple properties of LM residuals, but we can still look for egregious patterns

```
plot_model(rand.mod, type = 'diag')[[1]]
```



Shrinkage / partial pooling

Using random effects gives you a slightly different estimate than fixed effects

```
rand.mod = lmer(log(Cry_ero) ~ 1 + (1|WaterbodyID), data = crysub1)
fixed.mod = lm(log(Cry_ero) ~ WaterbodyID, data = crysub1)
```

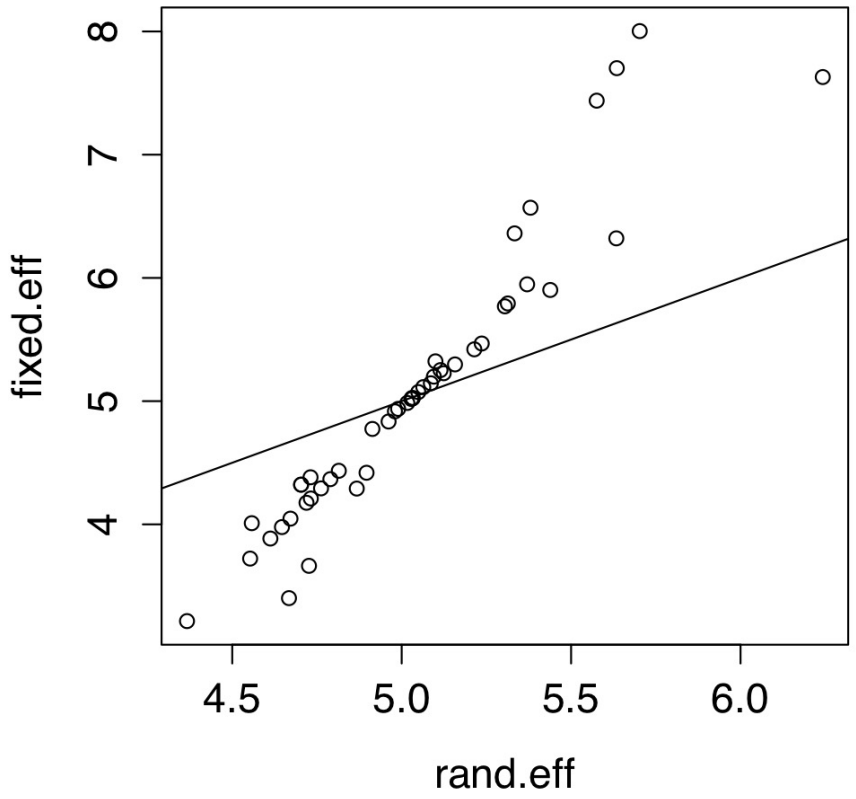
```
Call:
lm(formula = log(Cry_ero) ~ WaterbodyID, data = crysub1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0725 -0.4576  0.0000  0.3828  2.1684

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.17521    0.75564   5.525 1.07e-06 ***
WaterbodyID402  1.05126    0.97552   1.078 0.286170
WaterbodyID403  0.85056    0.97552   0.872 0.387267
WaterbodyID404  3.82781    1.30880   2.925 0.005099 **
WaterbodyID405 -0.19733    1.06863  -0.185 0.854219
WaterbodyID406  0.14644    0.97552   0.150 0.881259
WaterbodyID410 -0.16575    0.97552  -0.170 0.865737
WaterbodyID411  1.07666    1.06863   1.008 0.318353
WaterbodyID501  1.02350    1.06863   0.958 0.342614
WaterbodyID502  1.72685    0.97552   1.770 0.082560 .
WaterbodyID503  0.24363    1.30880   0.186 0.853056
```

```
fixed.effects = effect("WaterbodyID", fixed.mod)
fixed.eff = fixed.effects$fit[,1]
rand.eff = ranef(rand.mod)$WaterbodyID[,1] + fixef(rand.mod)[1]
```

```
plot(fixed.eff ~ rand.eff)
abline(a = 0, b = 1)
```



- In a fixed effects model, **only within-lake observations are used to estimate** that lake's parameter

- E.g. (4.5 + 4.7 + 5.1)/3 = 4.77

- For random effects, the **estimate depends on data from other lakes as well**

- Because the lakes are assumed to come from a common distribution, the best estimate for a lake depends on the parameters of that distribution

- If the across-lake mean is 5.04 and the Std.Dev is 0.59, then what happens if a lake has two observations, 2.4 and 4.6?

- It is very unlikely that the lake mean is 3.5. The model estimate essentially accounts for the fact that two observations will lead to a poor estimate

- The model estimate will be a **compromise between the lake mean and the overall mean:**

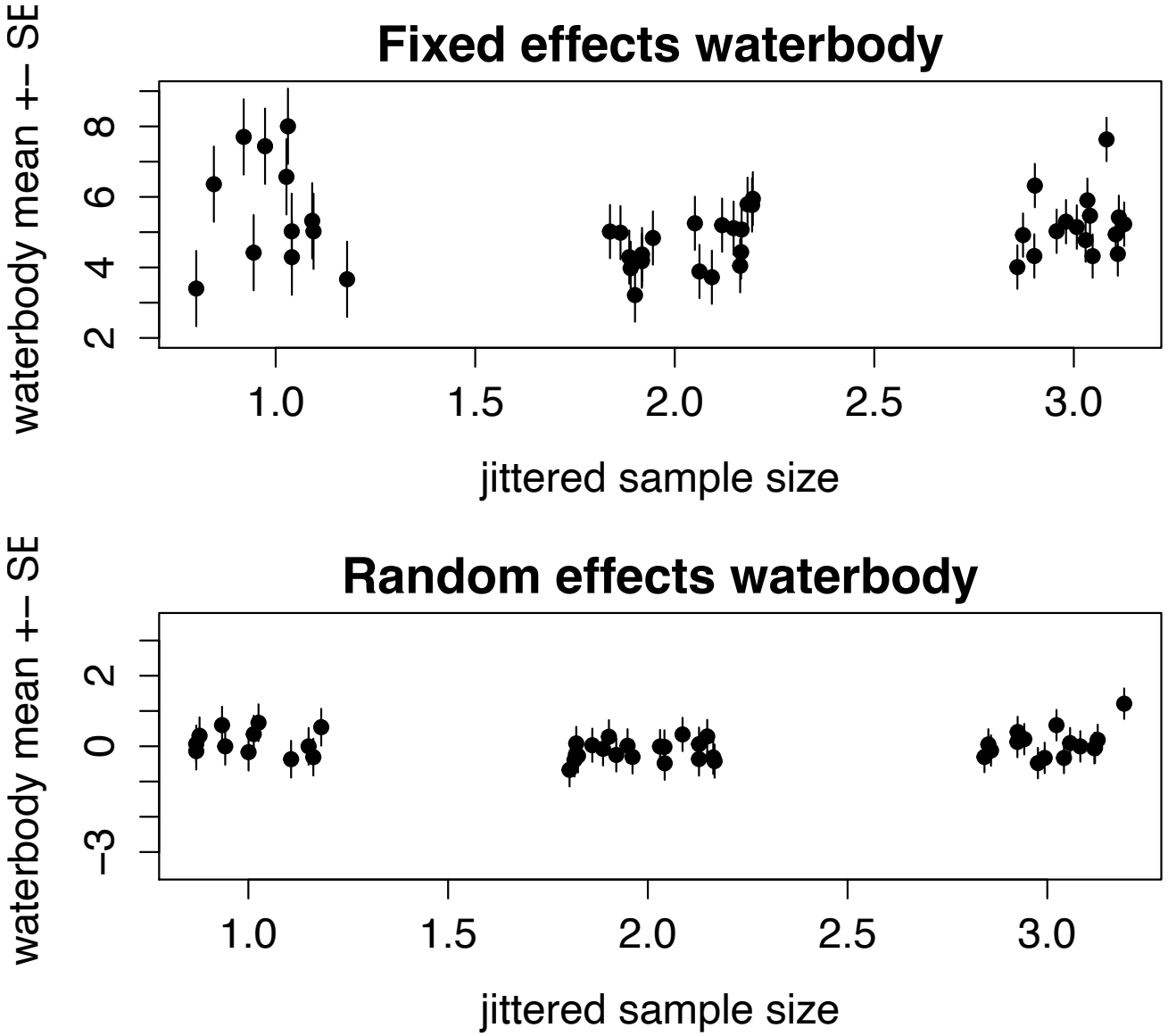
$$\hat{\alpha}_j \approx \frac{\frac{n_j}{\sigma_Y^2}\bar{Y}_j + \frac{1}{\sigma_\alpha^2}\bar{Y}_{all}}{\frac{n_j}{\sigma_Y^2} + \frac{1}{\sigma_\alpha^2}}$$

- This is called a shrinkage estimator. Amount of shrinkage depends on how many observations in that group

- Also referred to as ‘partial pooling’

- Seems weird that the estimated density in a lake should depend on what's happening in other lakes

- But this actually leads to more accurate results, if the assumptions are correct

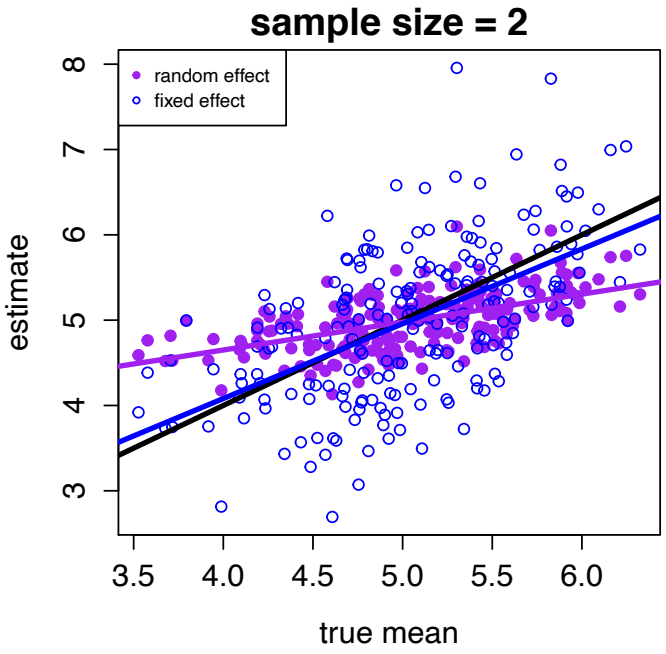


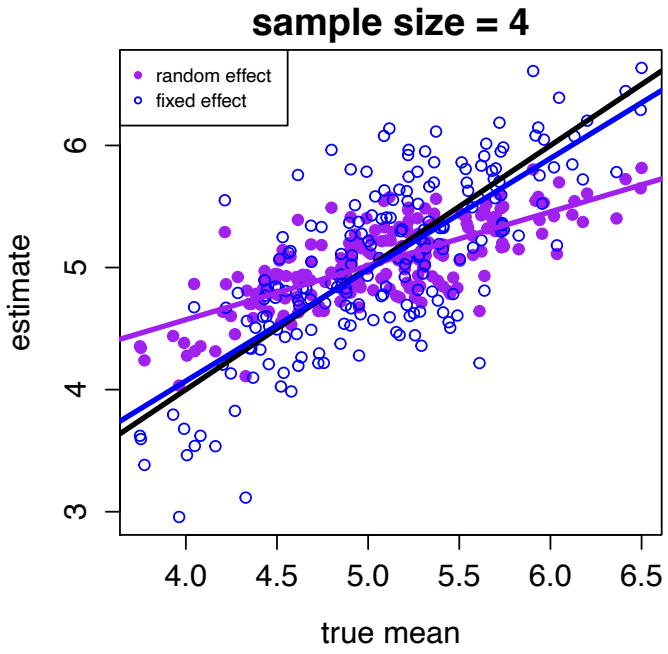
A shrinkage simulation

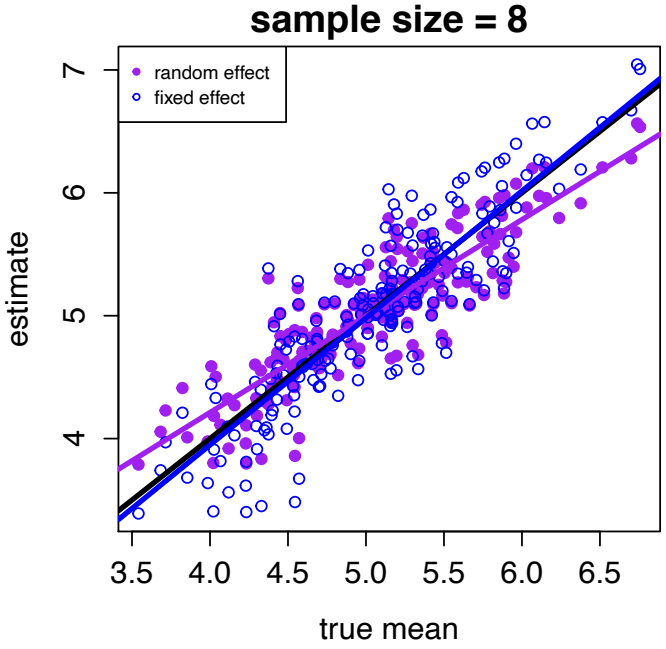
- 200 imaginary lakes; each lake gets a lake-specific mean
- Hyper-mean = 5, hyper-SD = 0.6
- each lake is sampled N times based on its lake-specific mean

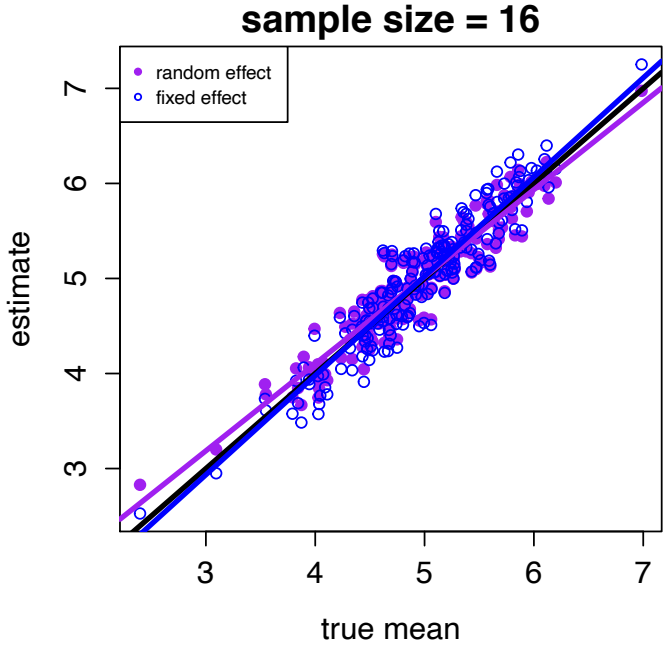
Compare fixed effects and random effects estimates

A shrinkage simulation









Bias vs. accuracy

Underfitting vs. overfitting