

Inference with mixed effects

Data and Goal

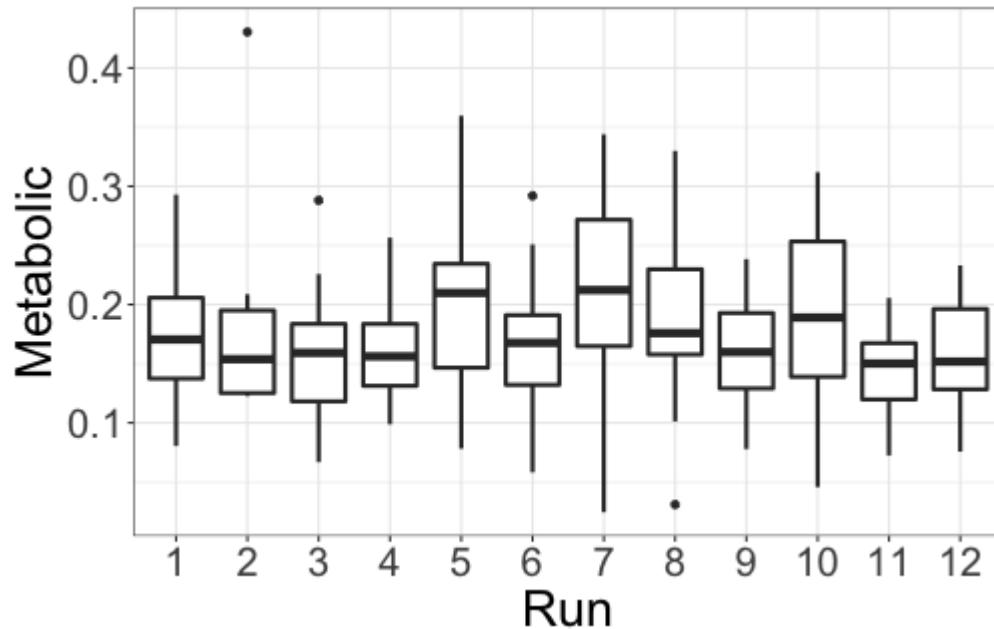
(marine bryozoans)

We have data on 197 early-stage *Bugula neritina*, with information on

- + Run: which repetition of the experiment the individual was recorded in
- + Mass: Mass of the individual (in micrograms)
- + Metabolic: Recorded metabolic rate (rate of energy consumption) of the individual (in mJ per hour)

Goal for this class: Is there systematic variation between different runs (i.e., is there any correlation due to Run)?

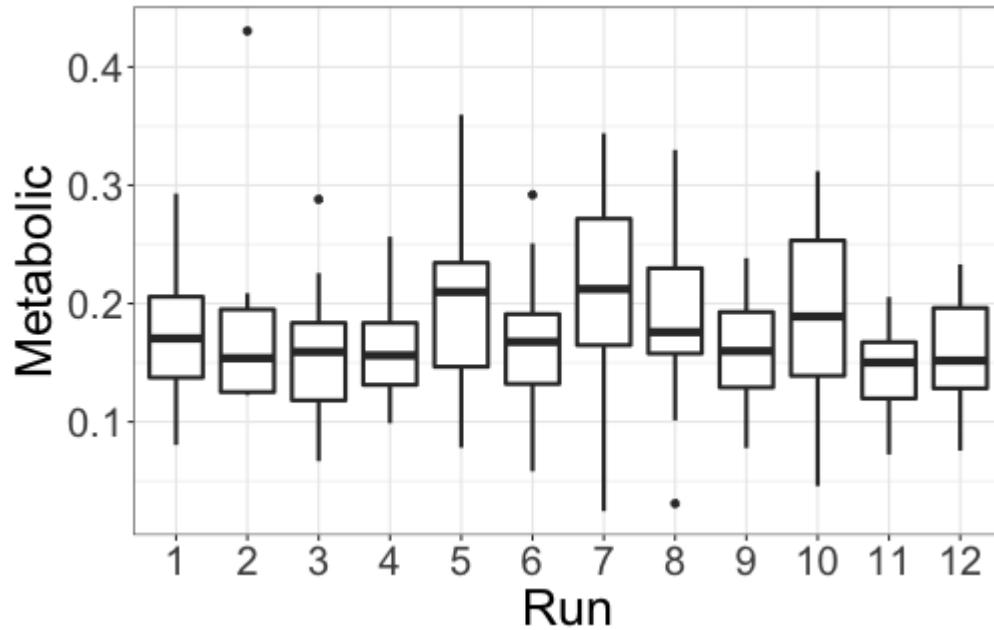
Visualizing runs



Does it look like there is variation from run to run (i.e., do we need u_i)?

- hard to say ; small differences in boxplot but nothing big
Differences could be due to random chance

Visualizing runs



It is hard to tell if there are actually differences between runs, or if observed differences are just due to chance.

How can I estimate variability between runs?

Initial random intercepts model

$$\text{Metabolic}_{ij} = \beta_0 + u_i + \varepsilon_{ij} \quad u_i \stackrel{\text{iid}}{\sim} N(0, \sigma_u^2), \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$$

random effect for run i

variability between runs

```
m1 <- lmer(Metabolic ~ (1|Run), data = bugula_early)
summary(m1)
```

random intercept
for each run

...

## Groups	Name	Variance	Std.Dev.	
## Run	(Intercept)	0.000131	0.01145	$\hat{\sigma}_u^2 = 0.000131$
## Residual		0.004181	0.06466	$\hat{\sigma}_\varepsilon^2 = 0.0042$
## Number of obs:	197, groups:	Run, 12		$\hat{\rho}_{\text{group}} = \frac{0.000131}{0.000131 + 0.0042}$
##				≈ 0.03
## Fixed effects:				
##	Estimate Std. Error t value			
## (Intercept)	0.174662	0.005692	30.68	
...				

About 3% of variation in metabolic rate is explained by differences between runs

Initial random intercepts model

$$Metabolic_{ij} = \beta_0 + u_i + \varepsilon_{ij} \quad u_i \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

$$\hat{\beta}_0 = 0.175, \quad \hat{\sigma}_u^2 = 0.00013, \quad \hat{\sigma}_\varepsilon^2 = 0.0042$$

How would we test whether there is systematic variation between runs?

$$H_0: \sigma_u^2 = 0$$

$$H_A: \sigma_u^2 > 0$$

$$\Rightarrow p_{group} = 0$$

$$\Rightarrow p_{group} > 0$$

Potential test statistic:

$$\hat{p}_{group} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2}$$

we should expect

\hat{p}_{group} to be close to 0 if H_0 is true,
but otherwise we don't actually know what the distribution is

Plan

$$H_0: \sigma_u^2 = 0$$

$$H_A: \sigma_u^2 > 0$$

$$\text{test stat.} = \hat{\rho}_{\text{group}}$$

p-value: requires us to approximate distribution of $\hat{\rho}_{\text{group}}$
if H_0 is true

- 1) Pretend H_0 is true \Rightarrow fit reduced model
e.g. $\text{Metabolic}_{ij} = \beta_0 + \varepsilon_{ij}$ (no u_i term)
- 2) Pretend reduced model is correct and simulate data from reduced model
- 3) Fit full model on simulated data and calculate $\hat{\rho}_{\text{group}}$
- 4) Repeat many times (e.g. w/ a for loop)

Intercept-only model (no random effect)

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

How do I fit this in R?

Intercept-only model (no random effect)

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

```
m0 <- lm(Metabolic ~ 1, data = bugula_early)
summary(m0)
```

...

```
##             Estimate Std. Error t value Pr(>|t|) 
## (Intercept) 0.174996   0.004672 37.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
## 
## Residual standard error: 0.06558 on 196 degrees of freedom
```

...

$$\hat{\beta}_0 = 0.175, \quad \hat{\sigma}_\varepsilon^2 = 0.06558^2 = 0.0043$$

Pretend the intercept-only model is correct

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \quad \underbrace{\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)}_{\text{captures randomness}}$$

I want to generate a new observation from this relationship.
What do I do?

• Sample new noise term $\varepsilon^* \sim N(0, \hat{\sigma}_\varepsilon^2)$

• Add ε^* to β_0

$$\text{Metabolic}^* = \hat{\beta}_0 + \varepsilon^*$$

Pretend the intercept-only model is correct

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

$$\hat{\beta}_0 = 0.175, \quad \hat{\sigma}_\varepsilon^2 = 0.0043$$

- + Sample a new noise term: $\varepsilon^* \sim N(0, 0.0043)$
- + Add the intercept: $0.175 + \varepsilon^*$

So our new observed metabolic rate is

$$Metabolic^* = 0.175 + \varepsilon^*$$

Simulation in R

Sample a new noise term: $\varepsilon^* \sim N(0, 0.0043)$

observations *mean* *standard deviation*

```
rnorm(n=1, mean=0, sd=sqrt(0.0043))
```

```
## [1] -0.06307819
```

Here $\varepsilon^* = -0.063$

Simulation in R

- + Sample a new noise term: $\varepsilon^* \sim N(0, 0.0043)$
- + Add the intercept: $0.175 + \varepsilon^*$

```
0.175 + rnorm(n=1, mean=0, sd=sqrt(0.0043))
```

```
## [1] 0.1119218
```

$$Metabolic^* = 0.112$$

In your R console, run the code to generate a new metabolic observation $Metabolic^*$.

Plan (so far)

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

How unusual are the observed differences between runs, if there is really no systematic differences between runs (i.e., no random effects)?

i) Fit reduced model ($\sigma_u^2 = 0$)

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij}$$

ii) Simulate from the reduced model

(So far, only simulated one observation)

Create a new dataset

Now we sample a new metabolic rate for **every** observation in the data:

$$Metabolic_{ij}^* = 0.175 + \varepsilon_{ij}^* \quad \varepsilon_{ij}^* \stackrel{iid}{\sim} N(0, 0.0043)$$

How can I modify my R code to sample a new metabolic rate for **every** observation?

change n
0.175 + rnorm(n=1, mean=0, sd=sqrt(0.0043))

Create a new dataset

Now we sample a new metabolic rate for **every** observation in the data:

$$Metabolic_{ij}^* = 0.175 + \varepsilon_{ij}^* \quad \varepsilon_{ij}^* \stackrel{iid}{\sim} N(0, 0.0043)$$

```
new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))
```

Create a new dataset

$$Metabolic_{ij}^* = 0.175 + \varepsilon_{ij}^* \quad \varepsilon_{ij}^* \stackrel{iid}{\sim} N(0, 0.0043)$$

```
new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))

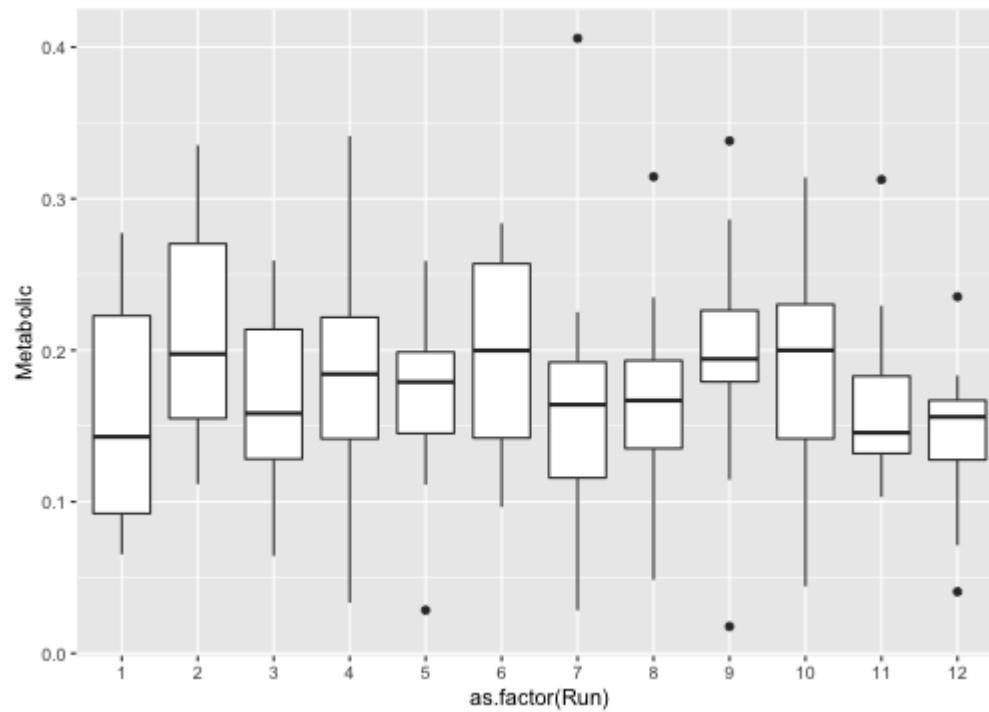
new_data <- data.frame(Run = bugula_early$Run,
                        Metabolic = new_metabolic)
```

- + Create a new metabolic rate for every organism in the data
- + Use the same runs from the observed data
- + Store the simulated dataset as new_data

Next: Fit the full model on the new data,
and calculate $\hat{\rho}_{group}$

Create a new dataset

```
new_data %>%
  ggplot(aes(x = as.factor(Run),
             y = Metabolic)) +
  geom_boxplot()
```



Plan (so far)

$$Metabolic_{ij} = \beta_0 + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

How unusual are the observed differences between runs, if there is really no systematic differences between runs (i.e., no random effects)?

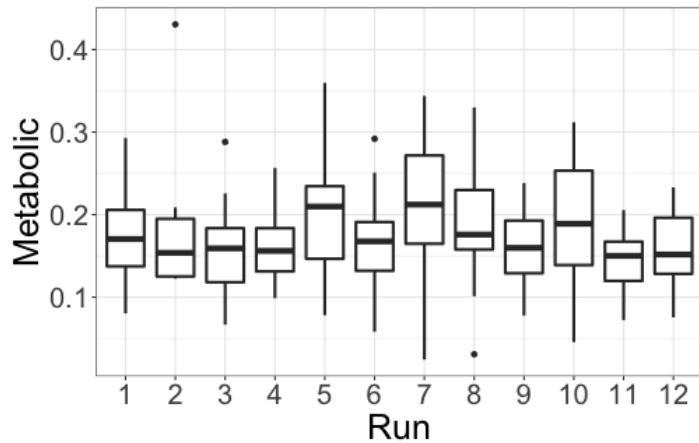
- + Pretend that the intercept-only model is correct
 - + $Metabolic^* = 0.175 + \varepsilon^* \quad \varepsilon^* \sim N(0, 0.0043)$
- + Create a new dataset from the intercept-only model

```
new_metabolic <- 0.175 +
  rnorm(n=197, mean=0, sd=sqrt(0.0043))
```

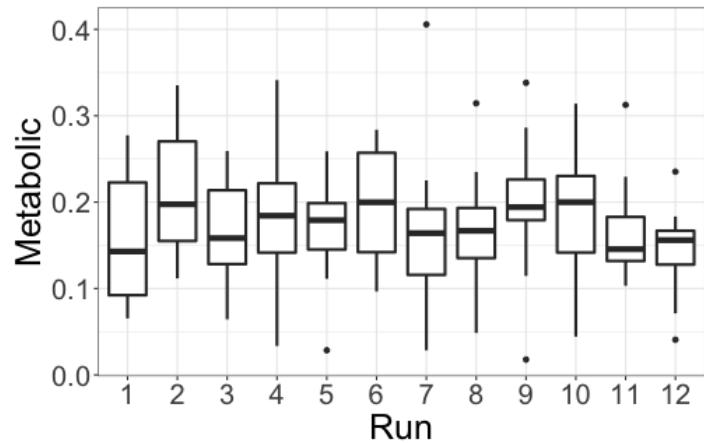
- + Compare our new dataset to the observed dataset

Compare new dataset to observed dataset

Original (observed) data:



New (simulated) data:



Do you think there is systematic variation between runs, or do you think the observed differences between runs are due to chance?

Compare new dataset to observed dataset

$$Metabolic_{ij} = \beta_0 + u_i + \varepsilon_{ij} \quad u_i \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

Fitted random intercepts model (observed data):

$$\hat{\beta}_0 = 0.175, \quad \hat{\sigma}_u^2 = 0.00013, \quad \hat{\sigma}_\varepsilon^2 = 0.0042 \quad \hat{P}_{grap} = 0.03$$

Fitted random intercepts model (simulated data):

$$\hat{\beta}_0 = 0.169, \quad \hat{\sigma}_u^2 = 0.00015, \quad \hat{\sigma}_\varepsilon^2 = 0.0049$$

$$\hat{P}_{grap} = \frac{0.00015}{0.000015 + 0.0049} \approx 0.03$$

Do you think there is systematic variation between runs, or do you think the observed differences between runs are due to chance?

Summary (so far)

Are there systematic differences between runs (group effects), or are observed differences simply due to chance?

- + Fit a model with no random effects
- + Simulate data from fitted model
- + Compare simulated data to observed data
 - + If simulated data looks very different, maybe there are systematic differences between runs
 - + If simulated data looks similar to observed data, maybe there aren't systematic differences between runs