Duke
UNIVERSITY

Linear Mixed Models

Data Club

Pol Castellano-Escuder, Ph.D. & **Akshay Bareja, Ph.D.**

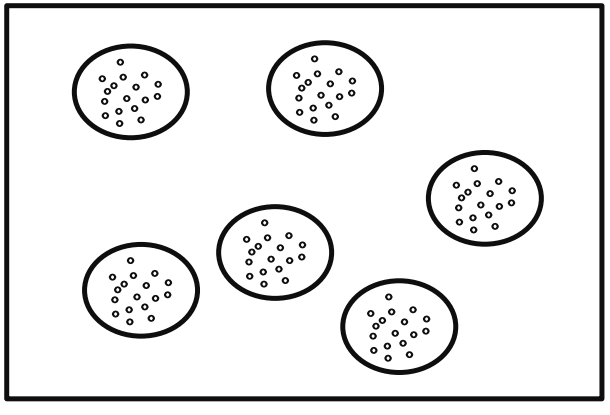
Duke Molecular Physiology Institute

May 5, 2022

What are Linear Mixed Models?

> Linear Mixed Models (LMMs) are an extension of linear models (LMs) to allow both fixed and random effects

LMMs are used when there is **non independence** in the data, the data have **multiple levels**, **hierarchical** structure, the data is **longitudinal**, and/or when the data are **correlated**. Therefore, data might have one or multiple sources of **clustering**, balanced or not.

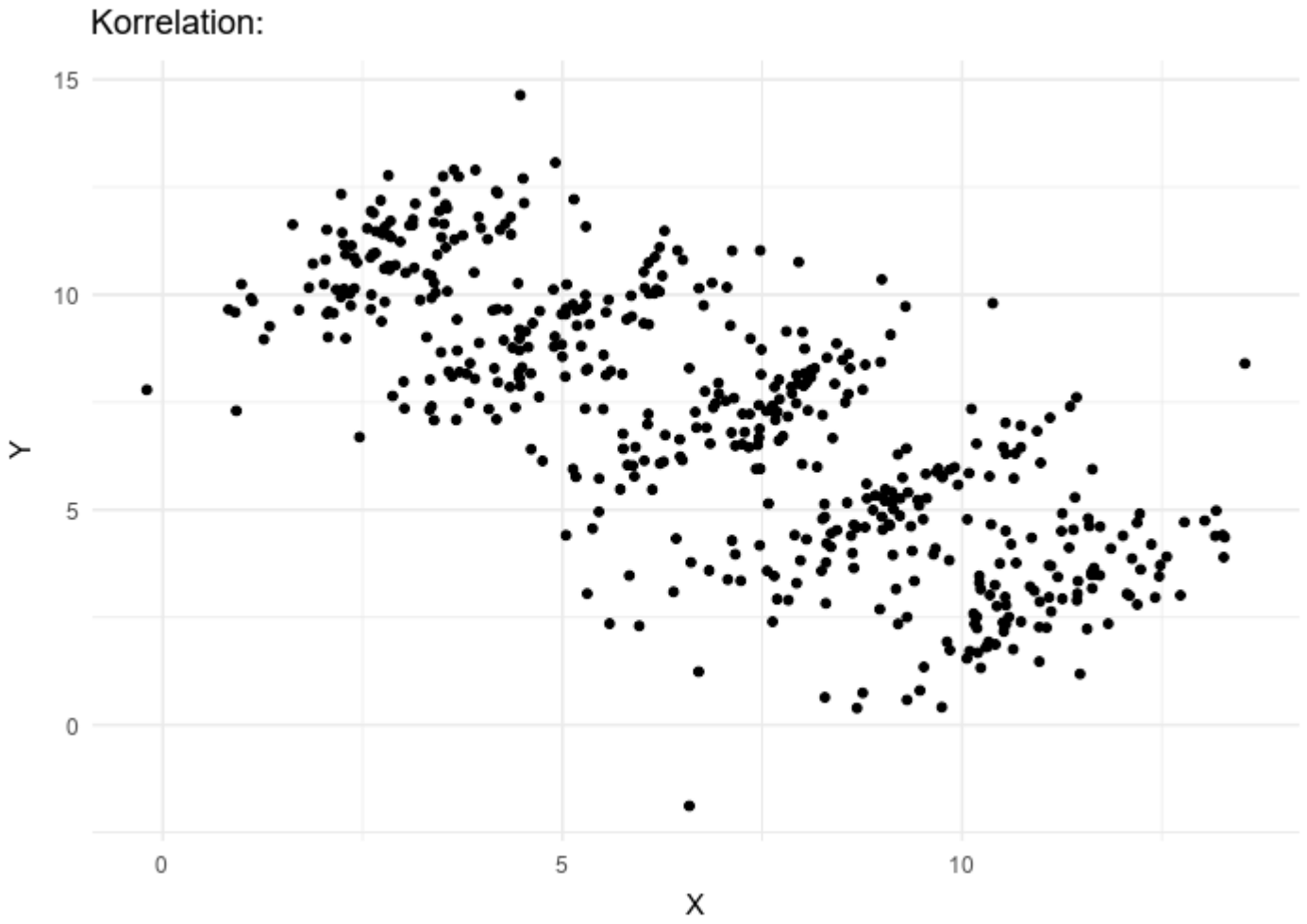


Option 1: Aggregate data ❌

Option 2: Multiple LMs by group ❌

Option 3: Use LMMs ✅

Simpson's Paradox



LMMs and ANOVA

- ANOVA: one continuous outcome variable and one or more categorical covariates
- LMMs: one continous outcome variable, one or more random effects and one or more fixed effects

When to fit a LMM?

- Continuous outcome variable
- Clustered data
- Sufficient number of clusters to enable estimation of the random effect (variance)

We should use LMMs instead of ANOVA when:

- Missing data
- Multiple (nested) random effects
- Not continous outcome (e.g., count and nominal data) -> GLMM

The `variancePartition` Package

Wrapper for **`lme4`** package, which is designed to fit LMMs

"Categorical variables should (almost) always be modeled as a random effect. The difference between modeling a categorical variable as a fixed versus random effect is minimal when the sample size is large compared to the number of categories (i.e. levels). So variables like disease status, sex or time point will not be sensitive to modeling as a fixed versus random effect. However, variables with many categories [...] must be modeled as a random effect in order to obtain statistically valid results. ***So to be on the safe side, categorical variable should be modeled as a random effect****."*

`variancePartition` fits two types of models:

1) Linear Mixed Model, where **all categorical variables are modeled as random effects** and **all continuous variables are fixed effects**. The function `lme4::lmer()` is used here.

2) Fixed effects model, where **all variables are modeled as fixed effects**. The function `lm` is used here.

Use Case

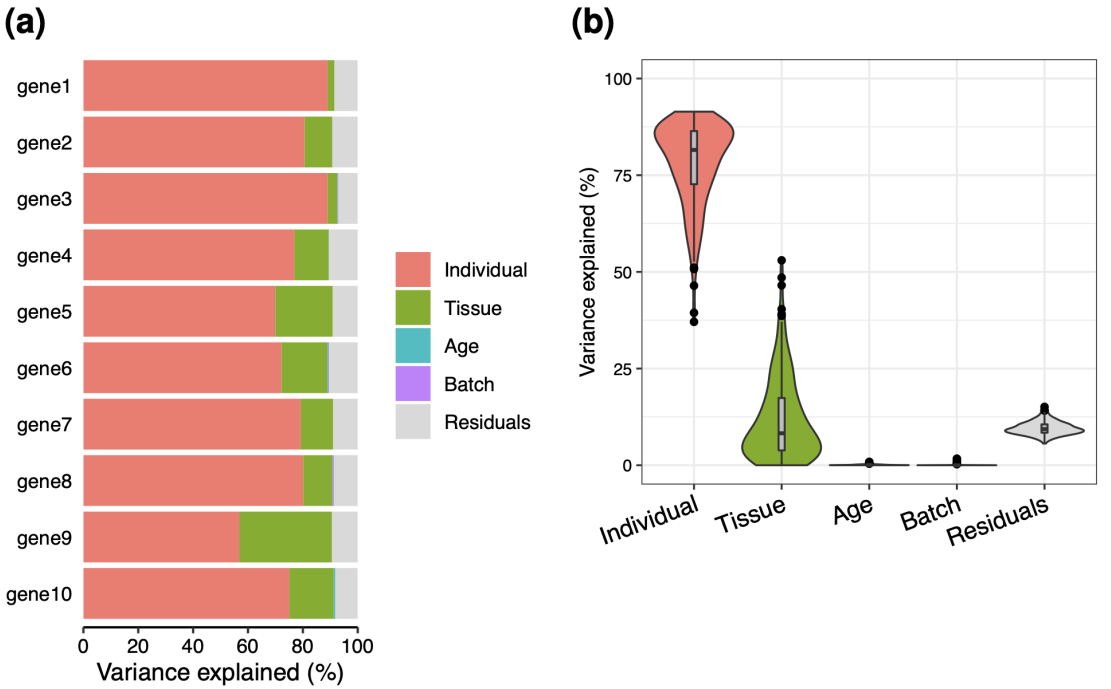
ID	Individual	Tissue	Batch	Age	Height
s1	1	A	1	44	157.4983
s2	5	C	4	42	196.6674
s3	9	C	2	49	189.4191
s4	13	A	3	48	156.2051
s5	18	A	1	56	178.4908
s6	22	B	4	43	175.4234
s7	26	A	1	37	150.4507
s8	30	A	4	48	194.0546
s9	34	B	1	50	193.9133
s10	38	B	2	56	157.0615

...

`Individual` has 25 levels, `Tissue` has 3 levels, and `Batch` has 4 levels

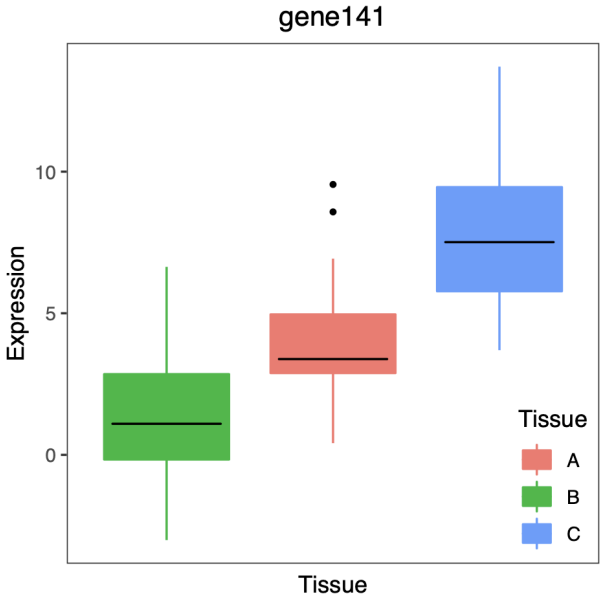
Design

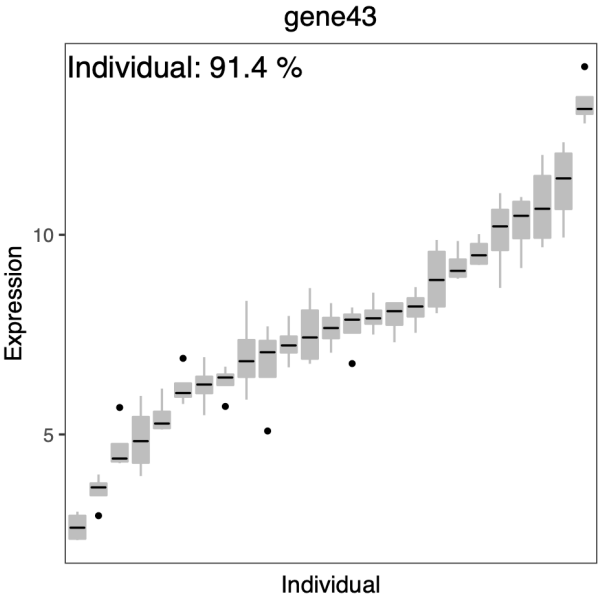
```
form <- ~ Age + (1|Individual) + (1|Tissue) + (1|Batch)

varPart <- fitExtractVarPartModel(geneExpr, form, info)
```



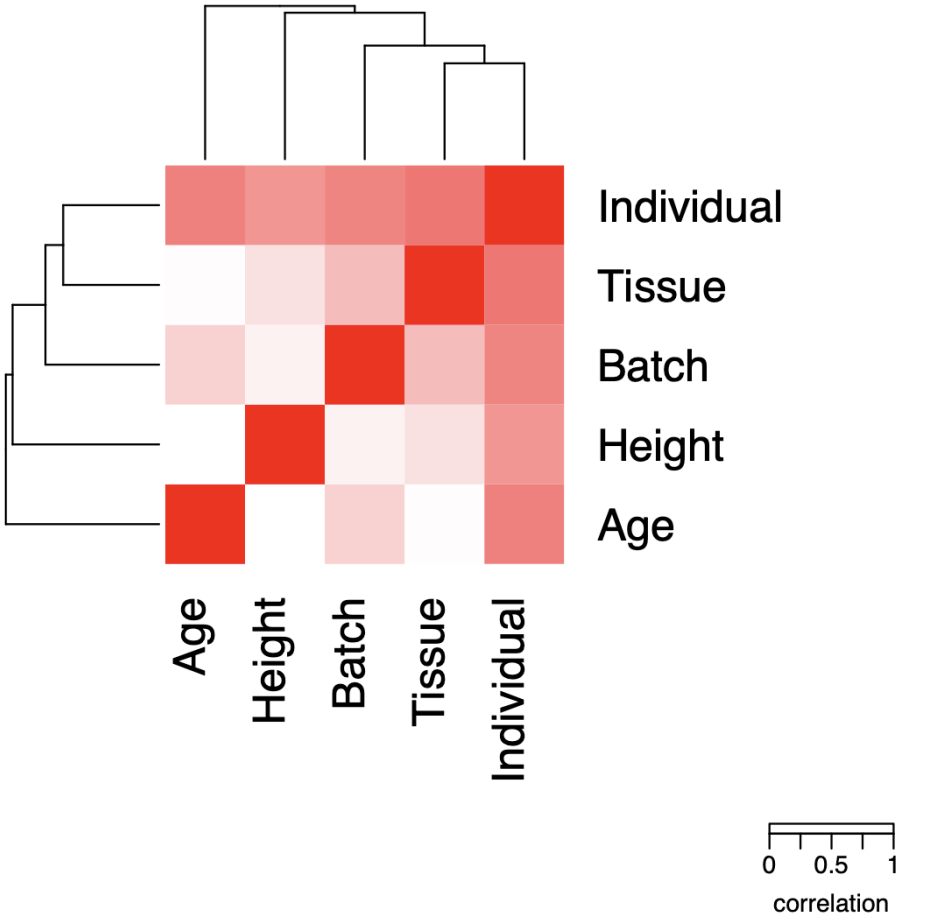
Plot Stratified Outcome

By Tissue





Correlations



Reproduce with `lme4`

```
library(lme4)

form_test <- geneExpr[1,] ~ Age + (1|Individual) + (1|Tissue)

fit <- lmer(form_test, info, REML = FALSE)

calcVarPart(fit)
```

```
>   Individual       Tissue          Age    Residuals
> 8.903140e-01 2.468013e-02 4.354738e-05 8.496235e-02
```