Linear models II

Steven Moran & Marco Maiolini

In the previous lecture...

Model	Variables	Distribution	R code
Linear Regression	$Y = b_0 + b_1x$	Normal	*lm(formula, data)*
General Linear Models	$Y = b_0 + b_1x_1 + b_2x_2 + \ldots$	Normal	*lm(formula, data)*
Generalized Linear Models (GLM)	$Y = b_0 + b_1x_1 + b_2x_2 + \ldots$	Any	*glm(formula, family, data)*

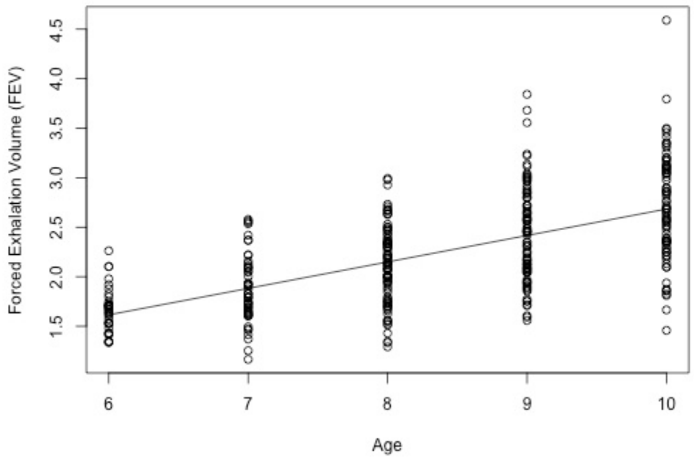
Linear Models (LMs)

In most models that we have seen (LM or GLM) we were interested in quantifying the exact effect of each explanatory variable.

Linear Models (LMs)

In most models that we have seen (LM or GLM) we were interested in quantifying the exact effect of each explanatory variable.

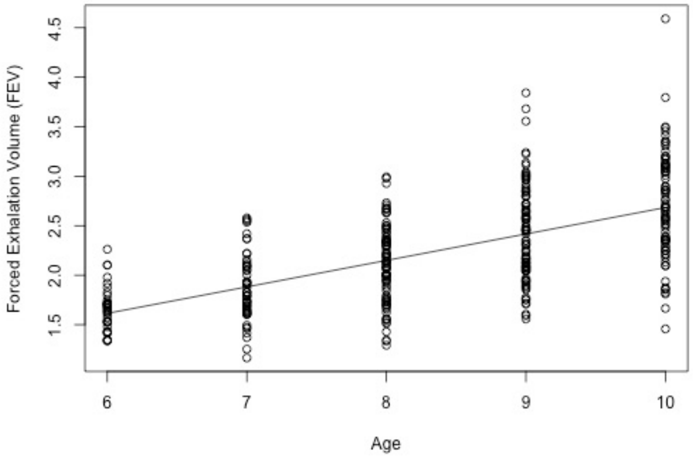
Differences between age of FEV

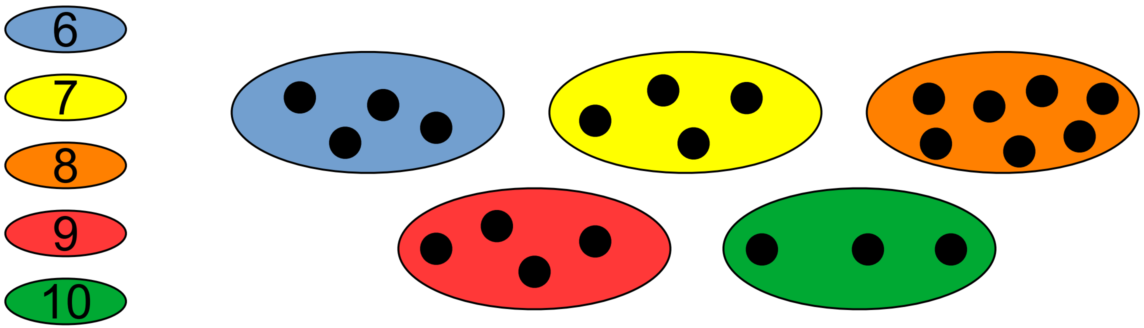


Linear Models (LMs)

In most models that we have seen (LM or GLM) we were interested in quantifying the exact effect of each explanatory variable.

Differences between age of FEV





● = Individual sampled

<u>The samples of each level are selected randomly, but the levels (Age) of X are not randomly chosen</u>

Linear Models (LMs)

In most models that we have seen (LM or GLM) we were interested in quantifying the exact effect of each explanatory variable.

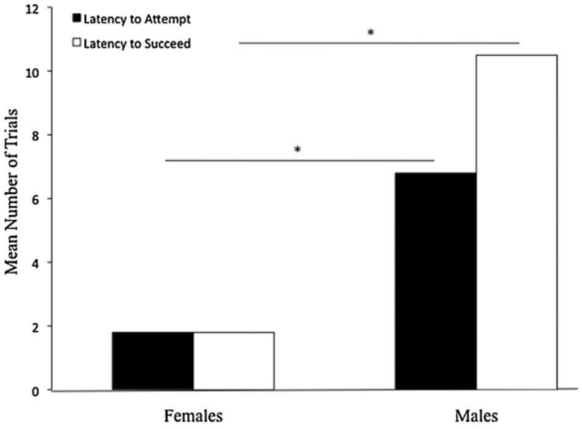
Differences between sex in tool use acquisition in Bonobos

Boose et al., 2013; “Sex differences in tool use acquisition in bonobos (pan paniscus)”, American Journal of Primatology.

Linear Models (LMs)

In most models that we have seen (LM or GLM) we were interested in quantifying the exact effect of each explanatory variable.

Differences between sex in tool use acquisition in Bonobos

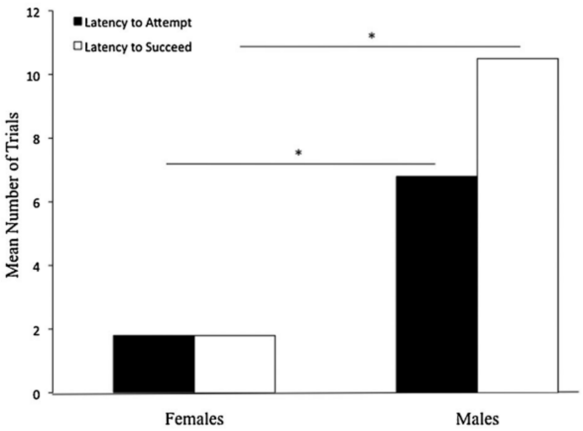


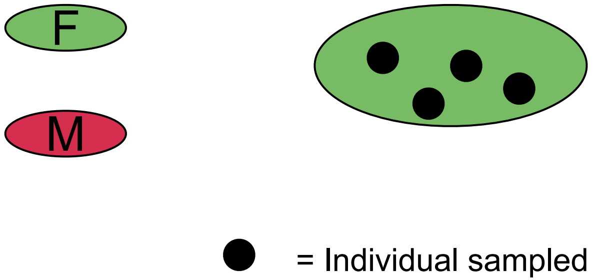
Boose et al., 2013; “Sex differences in tool use acquisition in bonobos (pan paniscus)”, American Journal of Primatology.

Linear Models (LMs)

In most models that we have seen (LM or GLM) we were interested in quantifying the exact effect of each explanatory variable.

Differences between sex in tool use acquisition in Bonobos





● = Individual sampled

<u>The samples of each level are selected randomly, but the levels (Sex) of X are not randomly chosen</u>

Boose et al., 2013; “Sex differences in tool use acquisition in bonobos (pan paniscus)”, American Journal of Primatology.

Linear Models (LMs)

The levels of our study (Age, Sex) are called FIXED EFFECT.

- Each level of a Fixed Factor is not derived randomly among infinite possibilities

- In an experiment, we look at all the possible levels of a Fixed Factor

- If we change the levels of a Fixed Factor, we have to change the hypothesis and repeat the experiment
    - E.g., if we change our fixed effect from sex to age in our Bonobos experiment, we also need to change our hypothesis

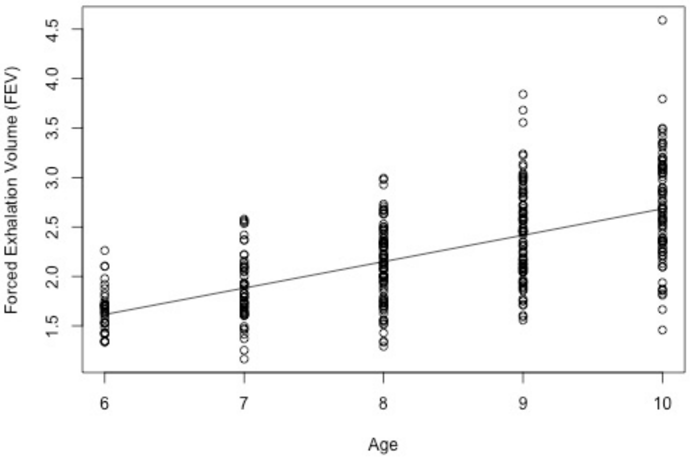
Linear Mixed Models (GLMMs)

Most of biological and behavioral data involves RANDOM EFFECTS, whose purpose is instead to quantify the variation among units.

Linear Mixed Models (GLMMs)

Most of biological and behavioral data involves RANDOM EFFECTS, whose purpose is instead to quantify the variation among units.

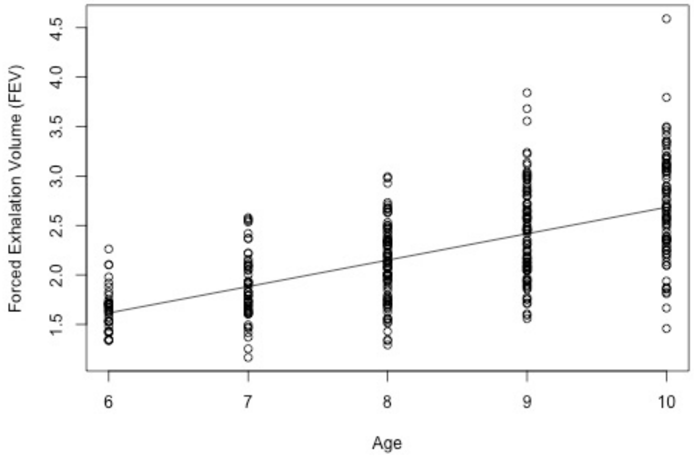
Differences between age of FAV, recorded in 2 schools.

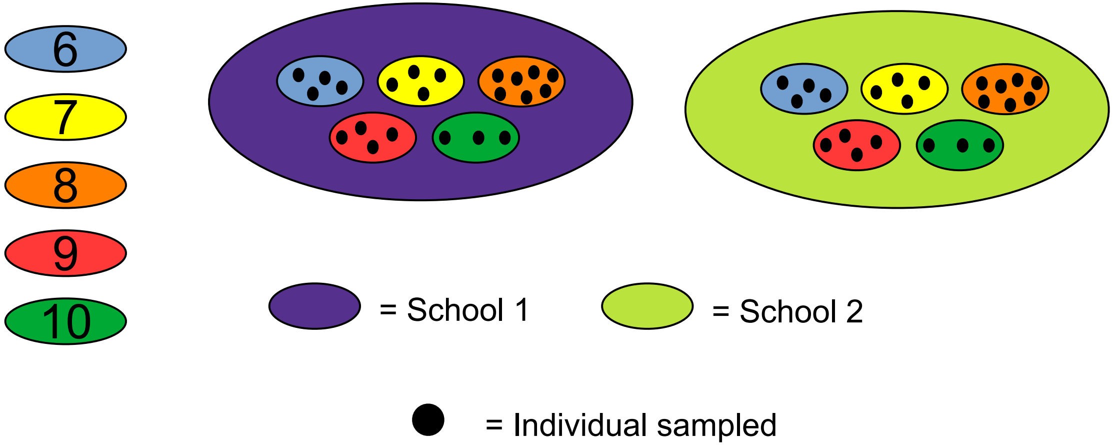


Linear Mixed Models (GLMMs)

Most of biological and behavioral data involves RANDOM EFFECTS, whose purpose is instead to quantify the variation among units.

Differences between age of FAV, recorded in 2 schools.





<u>The samples of each level are selected randomly within each school</u>

Linear Mixed Models (GLMMs)

Most of biological and behavioral data involves RANDOM EFFECTS, whose purpose is instead to quantify the variation among units.

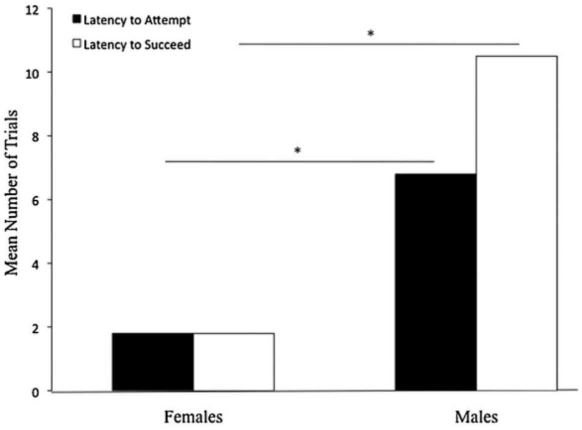
Differences between sex in tool use acquisition in Bonobos, in 3 different groups

Boose et al., 2013; “Sex differences in tool use acquisition in bonobos (pan paniscus)”, American Journal of Primatology.

Linear Mixed Models (GLMMs)

Most of biological and behavioral data involves RANDOM EFFECTS, whose purpose is instead to quantify the variation among units.

Differences between sex in tool use acquisition in Bonobos, in 3 different groups

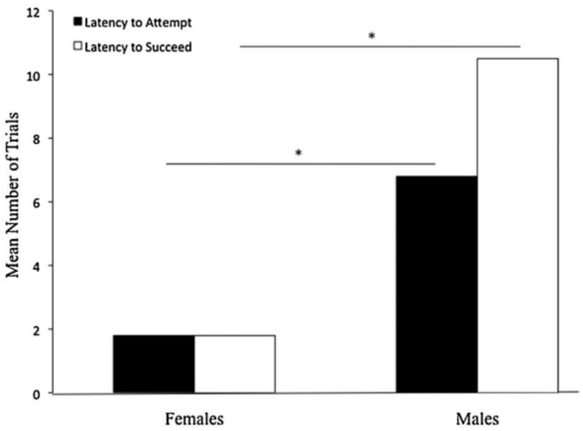


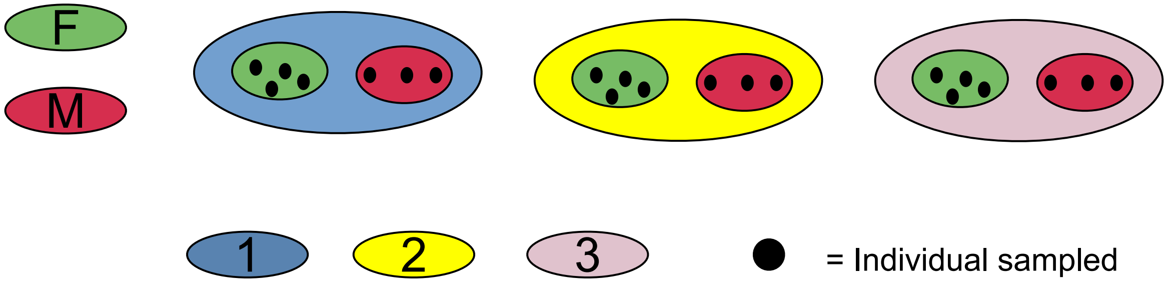
Boose et al., 2013; “Sex differences in tool use acquisition in bonobos (pan paniscus)”, American Journal of Primatology.

Linear Mixed Models (GLMMs)

Most of biological and behavioral data involves RANDOM EFFECTS, whose purpose is instead to quantify the variation among units.

Differences between sex in tool use acquisition in Bonobos, in 3 different groups





<u>The samples of each level are selected randomly within each group</u>

Boose et al., 2013; “Sex differences in tool use acquisition in bonobos (pan paniscus)”, American Journal of Primatology.

Linear Mixed Models (GLMMs)

The levels that quantify our variance are called RANDOM EFFECT.

- Each level of a Random Effect is collected randomly among infinite possibilities

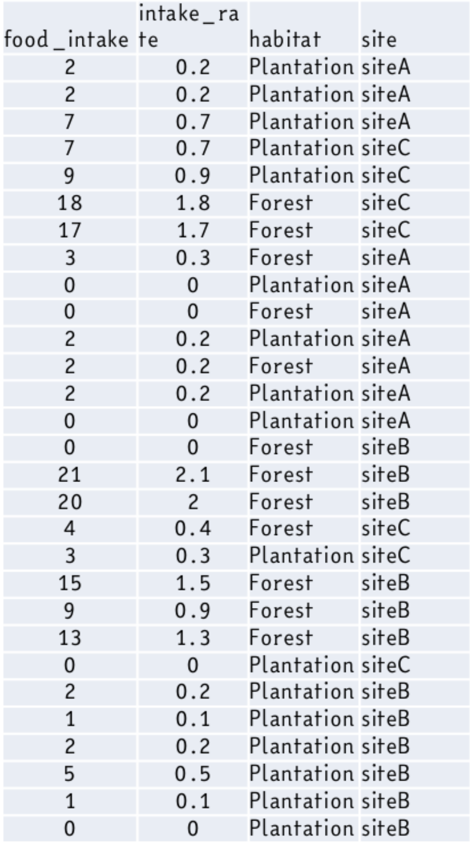
- In an experiment, we cannot take all the possible levels of a Random Effect

- If we change the levels of a Random Effect, we do not have to change our hypothesis.

Linear Mixed Models (LMMs)

Example

You want to measure how much is the food intake for a species in two habitat types (Forest and Plantation), measured in three different sites.



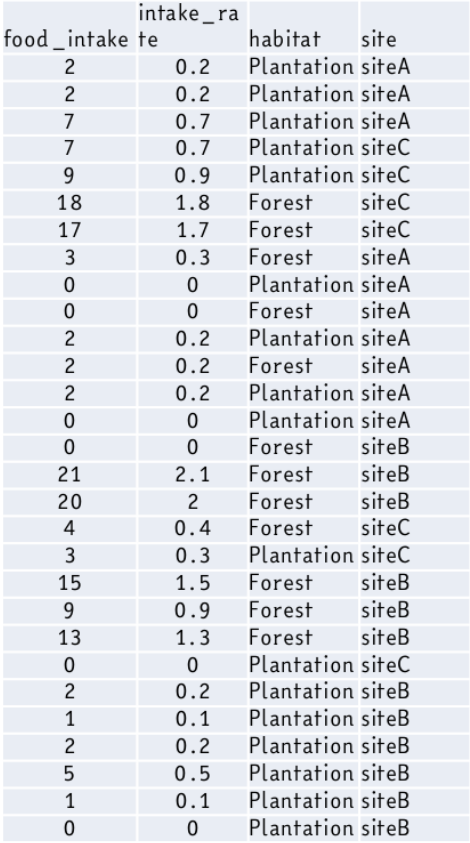
food_intake	intake_rate	habitat	site
2	0.2	Plantation	siteA
2	0.2	Plantation	siteA
7	0.7	Plantation	siteA
7	0.7	Plantation	siteC
9	0.9	Plantation	siteC
18	1.8	Forest	siteC
17	1.7	Forest	siteC
3	0.3	Forest	siteA
0	0	Plantation	siteA
0	0	Forest	siteA
2	0.2	Plantation	siteA
2	0.2	Forest	siteA
2	0.2	Plantation	siteA
0	0	Plantation	siteA
0	0	Forest	siteB
21	2.1	Forest	siteB
20	2	Forest	siteB
4	0.4	Forest	siteC
3	0.3	Plantation	siteC
15	1.5	Forest	siteB
9	0.9	Forest	siteB
13	1.3	Forest	siteB
0	0	Plantation	siteC
2	0.2	Plantation	siteB
1	0.1	Plantation	siteB
2	0.2	Plantation	siteB
5	0.5	Plantation	siteB
1	0.1	Plantation	siteB
0	0	Plantation	siteB

Linear Mixed Models (LMMs)

Example

You want to measure how much is the food intake for a species in two habitat types (Forest and Plantation), measured in three different sites.

Fixed Factor → Habitat



food_intake	intake_rate	habitat	site
2	0.2	Plantation	siteA
2	0.2	Plantation	siteA
7	0.7	Plantation	siteA
7	0.7	Plantation	siteC
9	0.9	Plantation	siteC
18	1.8	Forest	siteC
17	1.7	Forest	siteC
3	0.3	Forest	siteA
0	0	Plantation	siteA
0	0	Forest	siteA
2	0.2	Plantation	siteA
2	0.2	Forest	siteA
2	0.2	Plantation	siteA
0	0	Plantation	siteA
0	0	Forest	siteB
21	2.1	Forest	siteB
20	2	Forest	siteB
4	0.4	Forest	siteC
3	0.3	Plantation	siteC
15	1.5	Forest	siteB
9	0.9	Forest	siteB
13	1.3	Forest	siteB
0	0	Plantation	siteC
2	0.2	Plantation	siteB
1	0.1	Plantation	siteB
2	0.2	Plantation	siteB
5	0.5	Plantation	siteB
1	0.1	Plantation	siteB
0	0	Plantation	siteB

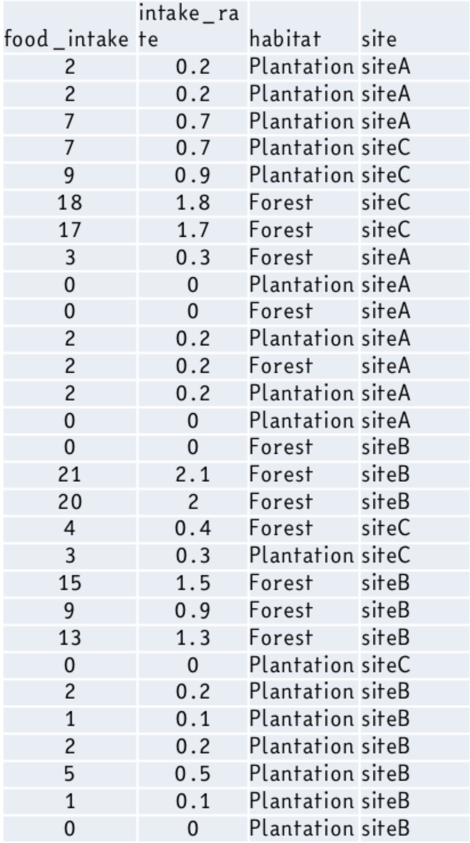
Linear Mixed Models (LMMs)

Example

You want to measure how much is the food intake for a species in two habitat types (Forest and Plantation), measured in three different sites.

Fixed Factor → Habitat

Random Effect → Sites



food_intake	intake_rate	habitat	site
2	0.2	Plantation	siteA
2	0.2	Plantation	siteA
7	0.7	Plantation	siteA
7	0.7	Plantation	siteC
9	0.9	Plantation	siteC
18	1.8	Forest	siteC
17	1.7	Forest	siteC
3	0.3	Forest	siteA
0	0	Plantation	siteA
0	0	Forest	siteA
2	0.2	Plantation	siteA
2	0.2	Forest	siteA
2	0.2	Plantation	siteA
0	0	Plantation	siteA
0	0	Forest	siteB
21	2.1	Forest	siteB
20	2	Forest	siteB
4	0.4	Forest	siteC
3	0.3	Plantation	siteC
15	1.5	Forest	siteB
9	0.9	Forest	siteB
13	1.3	Forest	siteB
0	0	Plantation	siteC
2	0.2	Plantation	siteB
1	0.1	Plantation	siteB
2	0.2	Plantation	siteB
5	0.5	Plantation	siteB
1	0.1	Plantation	siteB
0	0	Plantation	siteB

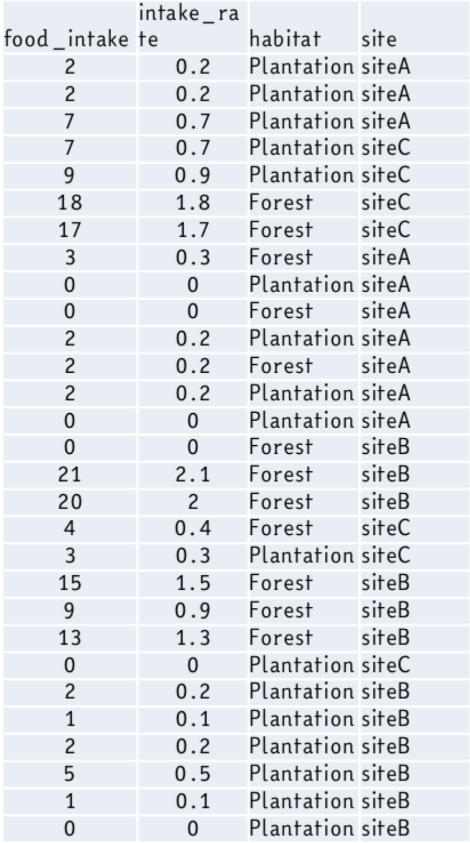
Linear Mixed Models (LMMs)

Example

You want to measure how much is the food intake for a species in two habitat types (Forest and Plantation), measured in three different sites.

Fixed Factor → Habitat

Random Effect → Sites



food_intake	intake_rate	habitat	site
2	0.2	Plantation	siteA
2	0.2	Plantation	siteA
7	0.7	Plantation	siteA
7	0.7	Plantation	siteC
9	0.9	Plantation	siteC
18	1.8	Forest	siteC
17	1.7	Forest	siteC
3	0.3	Forest	siteA
0	0	Plantation	siteA
0	0	Forest	siteA
2	0.2	Plantation	siteA
2	0.2	Forest	siteA
2	0.2	Plantation	siteA
0	0	Plantation	siteA
0	0	Forest	siteB
21	2.1	Forest	siteB
20	2	Forest	siteB
4	0.4	Forest	siteC
3	0.3	Plantation	siteC
15	1.5	Forest	siteB
9	0.9	Forest	siteB
13	1.3	Forest	siteB
0	0	Plantation	siteC
2	0.2	Plantation	siteB
1	0.1	Plantation	siteB
2	0.2	Plantation	siteB
5	0.5	Plantation	siteB
1	0.1	Plantation	siteB
0	0	Plantation	siteB

In R:

lmer(Response variable ~ Fixed Factor + (1|Random), data = dataset)

Linear Mixed Models (LMMs)

Example

You want to measure how much is the food intake for a species in two habitat types (Forest and Plantation), measured in three different sites.

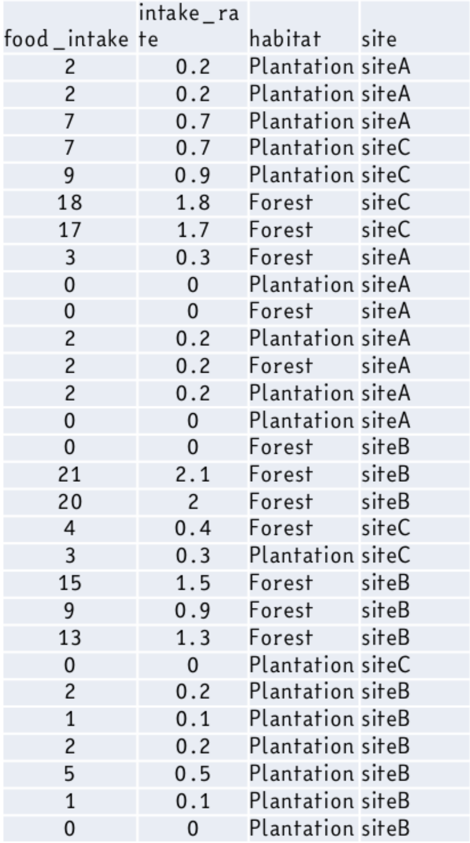
Fixed Factor → Habitat

Random Effect → Sites

lmer(intake_rate ~ habitat + (1|site), data = …)

In R:

lmer(Response variable ~ Fixed Factor + (1|Random), data = dataset)



food_intake	intake_rate	habitat	site
2	0.2	Plantation	siteA
2	0.2	Plantation	siteA
7	0.7	Plantation	siteA
7	0.7	Plantation	siteC
9	0.9	Plantation	siteC
18	1.8	Forest	siteC
17	1.7	Forest	siteC
3	0.3	Forest	siteA
0	0	Plantation	siteA
0	0	Forest	siteA
2	0.2	Plantation	siteA
2	0.2	Forest	siteA
2	0.2	Plantation	siteA
0	0	Plantation	siteA
0	0	Forest	siteB
21	2.1	Forest	siteB
20	2	Forest	siteB
4	0.4	Forest	siteC
3	0.3	Plantation	siteC
15	1.5	Forest	siteB
9	0.9	Forest	siteB
13	1.3	Forest	siteB
0	0	Plantation	siteC
2	0.2	Plantation	siteB
1	0.1	Plantation	siteB
2	0.2	Plantation	siteB
5	0.5	Plantation	siteB
1	0.1	Plantation	siteB
0	0	Plantation	siteB

Linear Mixed Models (LMMs)

Example

You want to measure how much is the food intake for a species in two habitat types (Forest and Plantation), measured in three different sites.

Fixed Factor → Habitat

Random Effect → Sites

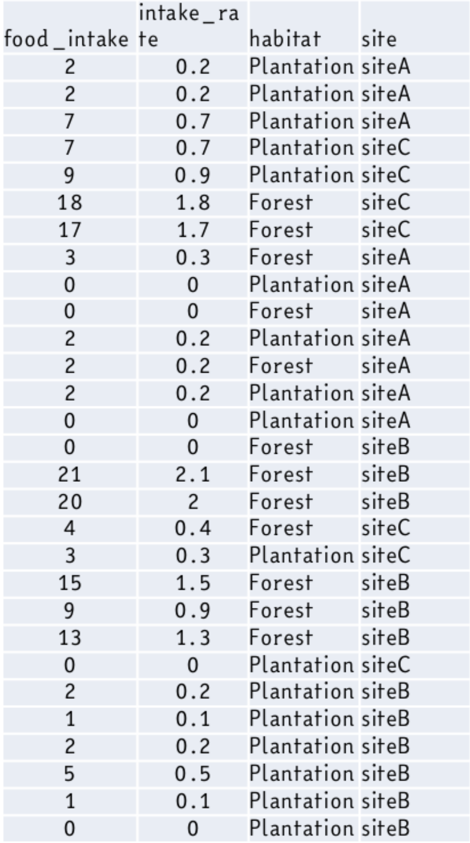
lmer(intake_rate ~ habitat + (1|site), data = …)

Remember to check if the residuals are normally distributed

A ← residuals(lmer(intake_rate ~ habitat + (1|site), data = …))
Shapiro.test(A)

In R:

lmer(Response variable ~ Fixed Factor + (1|Random), data = dataset)



food_intake	intake_rate	habitat	site
2	0.2	Plantation	siteA
2	0.2	Plantation	siteA
7	0.7	Plantation	siteA
7	0.7	Plantation	siteC
9	0.9	Plantation	siteC
18	1.8	Forest	siteC
17	1.7	Forest	siteC
3	0.3	Forest	siteA
0	0	Plantation	siteA
0	0	Forest	siteA
2	0.2	Plantation	siteA
2	0.2	Forest	siteA
2	0.2	Plantation	siteA
0	0	Plantation	siteA
0	0	Forest	siteB
21	2.1	Forest	siteB
20	2	Forest	siteB
4	0.4	Forest	siteC
3	0.3	Plantation	siteC
15	1.5	Forest	siteB
9	0.9	Forest	siteB
13	1.3	Forest	siteB
0	0	Plantation	siteC
2	0.2	Plantation	siteB
1	0.1	Plantation	siteB
2	0.2	Plantation	siteB
5	0.5	Plantation	siteB
1	0.1	Plantation	siteB
0	0	Plantation	siteB

Generalized Linear Mixed Models (GLMMs)

GLMMs combine the proprieties of Linear Mixed Models (LMM) and Generalized Linear Models (GLM), which handle non-normal data.

- You have a Random Effect

- It can handle data with poisson, binomial or normal distribution

- It can handle interactions and additive effects

Generalized Linear Mixed Models (GLMMs)

Example

You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Generalized Linear Mixed Models (GLMMs)

Example

You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Response → Rhythm

Generalized Linear Mixed Models (GLMMs)

Example

You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Response → Rhythm

Fixed Factor → Context

Generalized Linear Mixed Models (GLMMs)

Example

You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Response → Rhythm

Fixed Factor → Context

Random Effect → Individuals, groups and songs

Generalized Linear Mixed Models (GLMMs)

Example

You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Response → Rhythm

Fixed Factor → Context

Random Effect → Individuals, groups and songs

In R:

```
lmer(Response variable ~ Fixed Factor + (1|RE1) + (1|RE2), family = …, data = dataset)
```

Generalized Linear Mixed Models (GLMMs)

Example

You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Response → Rhythm

Fixed Factor → Context

Random Effect → Individuals, groups and songs

```
lmer(rhythm ~ context + (1|individual) + (1|group) + (1|song), family = …,
data = … )
```

In R:

```
lmer(Response variable ~ Fixed Factor + (1|RE1) + (1|RE2), family = …, data = dataset)
```

Generalized Linear Mixed Models (GLMMs)

Example

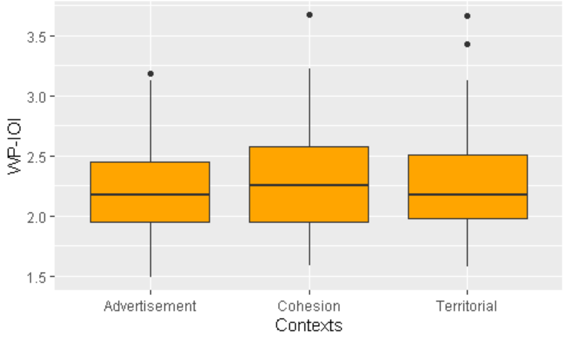
You want to check if there is a rhythmic difference among context in the indris' song.
You take the data from different individuals, groups and songs.

Response → Rhythm

Fixed Factor → Context

Random Effect → Individuals, groups and songs

lmer(rhythm ~ context + (1|individual) + (1|group) + (1|song), family = …, data = …)



In R:

lmer(Response variable ~ Fixed Factor + (1|RE1) + (1|RE2), family = …, data = dataset)

Model Selection

Ockham’s razor

Model Selection

Ockham’s razor → *“Pluralitas non est ponenda sine neccesitate”*

Between two theories that have same predictions, the simpler the better.

Model Selection

Ockham’s razor → *“Pluralitas non est ponenda sine neccesitate”*

Between two theories that have same predictions, <u>the simpler the better.</u>

When comparing different models we want to find out which model explain better my data regardless the **Individual independent variables** in the model.

❗ It doesn’t mean that there is a right and a wrong model!

Model Selection

Ockham's razor → *"Pluralitas non est ponenda sine neccesitate"*

Between two theories that have same predictions, the simpler the better.

When comparing different models we want to find out which model explain better my data regardless the **Individual independent variables** in the model.

❗ It doesn't mean that there is a right and a wrong model!

Akaike's information criteria (AIC):

AIC estimates model complexity. It works estimating the expected performance of model's predictions, for that scope it use observed data and hypothetical sample generated by the same model.

The best model show the smallest value; a difference within **4 - 7** units indicate less support, a difference over **10** indicate that the worse model can be omitted

Model Selection

Ockham's razor → *"Pluralitas non est ponenda sine neccesitate"*

Between two theories that have same predictions, the simpler the better.

When comparing different models we want to find out which model explain better my data regardless the **Individual independent variables** in the model.

! It doesn't mean that there is a right and a wrong model!

Akaike's information criteria (AIC):

AIC estimates model complexity. It works estimating the expected performance of model's predictions, for that scope it use observed data and hypothetical sample generated by the same model.

The best model show the smallest value; a difference within **4 - 7** units indicate less support, a difference over **10** indicate that the worse model can be omitted

In R:

AIC(model)

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

```
Model1 ← lmer(water_infection ~ river_tract + (1|id_rivers), data = ... )

Summary(Model1)

AIC(Model1)
```

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

```
Model1 ← lmer(water_infection ~ river_tract + (1|id_rivers), data = ... )

Summary(Model1)

AIC(Model1)
```

```
AIC BIC logLik deviance df. Resid

4838 4920 -2410 4820 900
```

AIC = 4838

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

```
Model1 ← lmer(water_infection ~ river_tract + (1|id_rivers), data = ... )
Summary(Model1)
AIC(Model1)
```

```
AIC BIC logLik deviance df. Resid
4838 4920 -2410 4820 900
```

AIC = 4838

```
Model2 ← lmer(water_infection ~ river_tract + slope + (1|id_rivers), data = ... )
Summary(Model2)
AIC(Model2)
```

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

```
Model1 ← lmer(water_infection ~ river_tract + (1|id_rivers), data = ... )

Summary(Model1)

AIC(Model1)
```

```
AIC BIC logLik deviance df. Resid

4838 4920 -2410 4820 900
```

AIC = 4838

```
Model2 ← lmer(water_infection ~ river_tract + slope + (1|id_rivers), data = ... )

Summary(Model2)

AIC(Model2)
```

```
AIC BIC logLik deviance df. Resid

4810 4911 -2287 4780 899
```

AIC = 4810

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

```
Model1 ← lmer(water_infection ~ river_tract + (1|id_rivers), data = ... )
Summary(Model1)
AIC(Model1)
```

```
AIC BIC logLik deviance df. Resid
4838 4920 -2410 4820 900
```

AIC = 4838

```
Model2 ← lmer(water_infection ~ river_tract + slope + (1|id_rivers), data = ... )
Summary(Model2)
AIC(Model2)
```

```
AIC BIC logLik deviance df. Resid
4810 4911 -2287 4780 899
```

4838 – 4810 = 28

AIC = 4810

Model Selection

Example

You want to check if there is a difference in water infection difference depending on the river tract sampled.
You sample different rivers.

```
Model1 ← lmer(water_infection ~ river_tract + (1|id_rivers), data = ... )
Summary(Model1)
AIC(Model1)
```

```
AIC BIC logLik deviance df. Resid
4838 4920 -2410 4820 900
```

AIC = 4838

```
Model2 ← lmer(water_infection ~ river_tract + slope + (1|id_rivers), data = ... )
Summary(Model2)
AIC(Model2)
```

```
AIC BIC logLik deviance df. Resid
4810 4911 -2287 4780 899
```

AIC = 4810

4838 – 4810 = 28

Model2 is better

An example

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa (Atkinson 2011)

- “Here I show that the number of phonemes used in a global sample of 504 languages is also clinal and fits a serial founder–effect model of expansion from an inferred origin in Africa.”

- Assumption: “The more speakers a language has, the bigger its phoneme inventory is likely to be.” (Hay & Bauer 2007)

An example

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa (Atkinson 2011)

- “Here I show that the number of phonemes used in a global sample of 504 languages is also clinal and fits a serial founder–effect model of expansion from an inferred origin in Africa.”

- Assumption: “The more speakers a language has, the bigger its phoneme inventory is likely to be.” (Hay & Bauer 2007)

- “For maximum comparability to earlier studies, we created models in which log(population) is the independent variable (more commonly called a fixed-effect predictor in the literature on mixed models). Language genus and language family were included as group-level predictors (a.k.a. random effects).” (Moran et al., 2012)