Statistical testing

Session 3 LOVE'R course

Pablo Raguet

email: pablo.raguet@inrae.fr
github: Capra-Ibex/R-course-2022

original teacher: Tania L. Maxwell
website: tania-maxwell.github.io

14-11-2022

Load the packages libraries

```
library(tidyverse)
library(multcomp)
library(MuMIn)
library(lme4)
library(nlme)

# not necesary in R project
# dir <- getwd()
# setwd(dir)
```

Topics for today

Generalized Linear Models (GLM)

- Link and Variance functions
- Deviance
- Over-dispersion

Topics for today

Generalized Linear Models (GLM)

- Link and Variance functions
- Deviance
- Over-dispersion

Linear Mixt Model (LMM)

- Fixed vs. random effects
- Pseudoreplication
- Linear mixed effect models (LMMs)

Topics for today

Generalized Linear Models (GLM)

- Link and Variance functions
- Deviance
- Over-dispersion

Linear Mixt Model (LMM)

- Fixed vs. random effects
- Pseudoreplication
- Linear mixed effect models (LMMs)

Heritability

Generalized Linear Models

Dataset for GLM

Import the `basketcuiller.csv` dataset:

```
Baskspo <- read.csv2("Data/Exemple/basketcuiller.csv") %>%
  mutate_if(is.character, as.factor) %>%
  mutate(dist = as_factor(dist))
```

Dataset for GLM

Import the `basketcuiller.csv` dataset:

```
Baskspo <- read.csv2("Data/Exemple/basketcuiller.csv") %>%
  mutate_if(is.character, as.factor) %>%
  mutate(dist = as_factor(dist))
```

Testing the ability to mark a bullet paper in trashcan according to:

- **Distance** (6, 8, 10 or 12 tiles)
- **Method** (hand or spoon)

Dataset for GLM

Import the `basketcuiller.csv` dataset:

```
Baskspo <- read.csv2("Data/Exemple/basketcuiller.csv") %>%
  mutate_if(is.character, as.factor) %>%
  mutate(dist = as_factor(dist))
```

Testing the ability to mark a bullet paper in trashcan according to:

- **Distance** (6, 8, 10 or 12 tiles)
- **Method** (hand or spoon)

Result :

- Success/Failure

Dataset for GLM

Import the `basketcuiller.csv` dataset:

```
Baskspo <- read.csv2("Data/Exemple/basketcuiller.csv") %>%
  mutate_if(is.character, as.factor) %>%
  mutate(dist = as_factor(dist))
```

Testing the ability to mark a bullet paper in trashcan according to:

- **Distance** (6, 8, 10 or 12 tiles)
- **Method** (hand or spoon)

Result :

- Success/Failure

With **4** distances, **2** methods, **4** group, **2** shooters per group and **4** trials per shooter: $4 \times 2 \times 4 \times 2 \times 4 = 256$ evaluations.

Dataset for GLM

Import the `basketcuiller.csv` dataset:

```
Baskspo <- read.csv2("Data/Exemple/basketcuiller.csv") %>%
  mutate_if(is.character, as.factor) %>%
  mutate(dist = as_factor(dist))
```

Testing the ability to mark a bullet paper in trashcan according to:

- **Distance** (6, 8, 10 or 12 tiles)
- **Method** (hand or spoon)

Result :

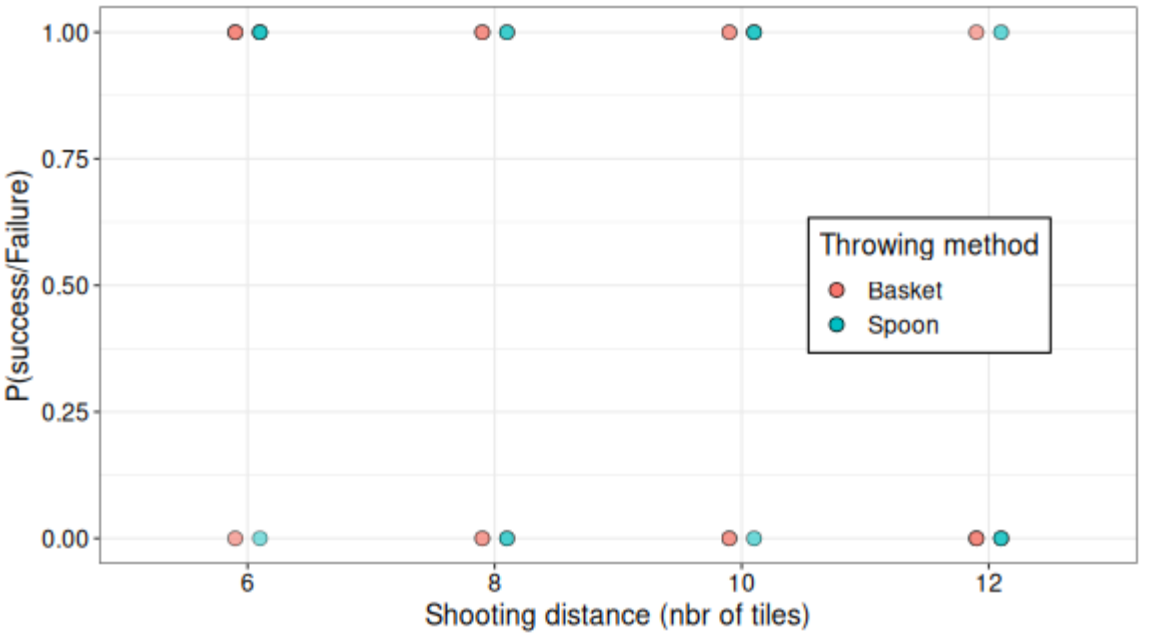
- Success/Failure

With **4** distances, **2** methods, **4** group, **2** shooters per group and **4** trials per shooter: $4 \times 2 \times 4 \times 2 \times 4 = 256$ evaluations.

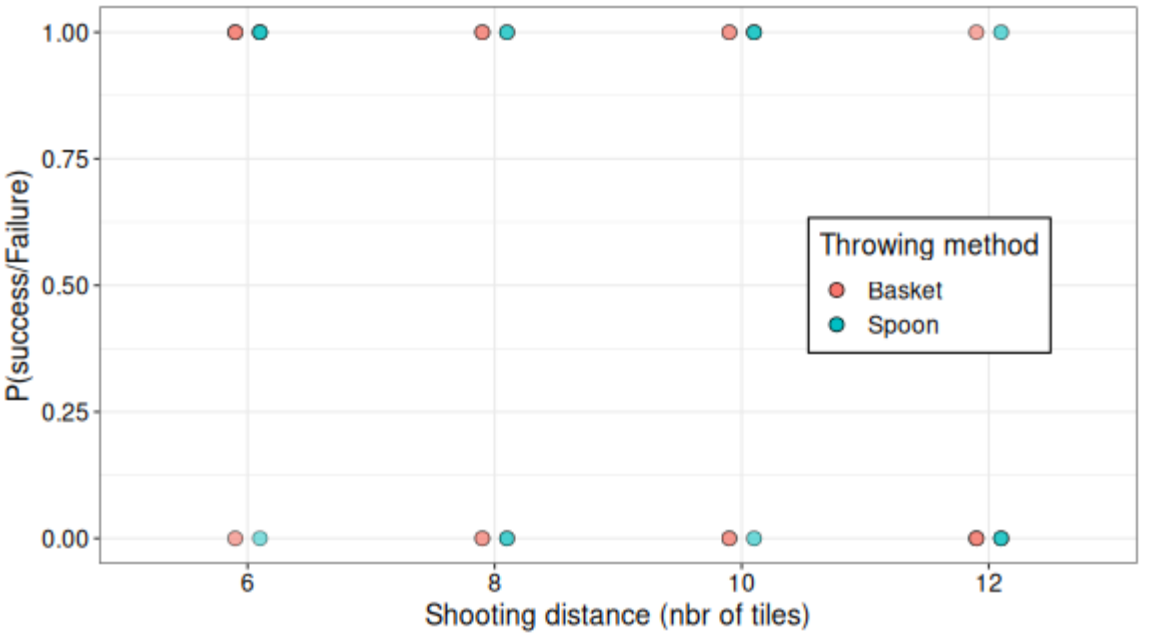
Thanks to Sebastien Ibanez and the students from ECOMONT master at Savoie Mont-Blanc University (2017).

Question: What is the influence of distance (and method) on succes ?

Question: What is the influence of distance (and method) on succes ?

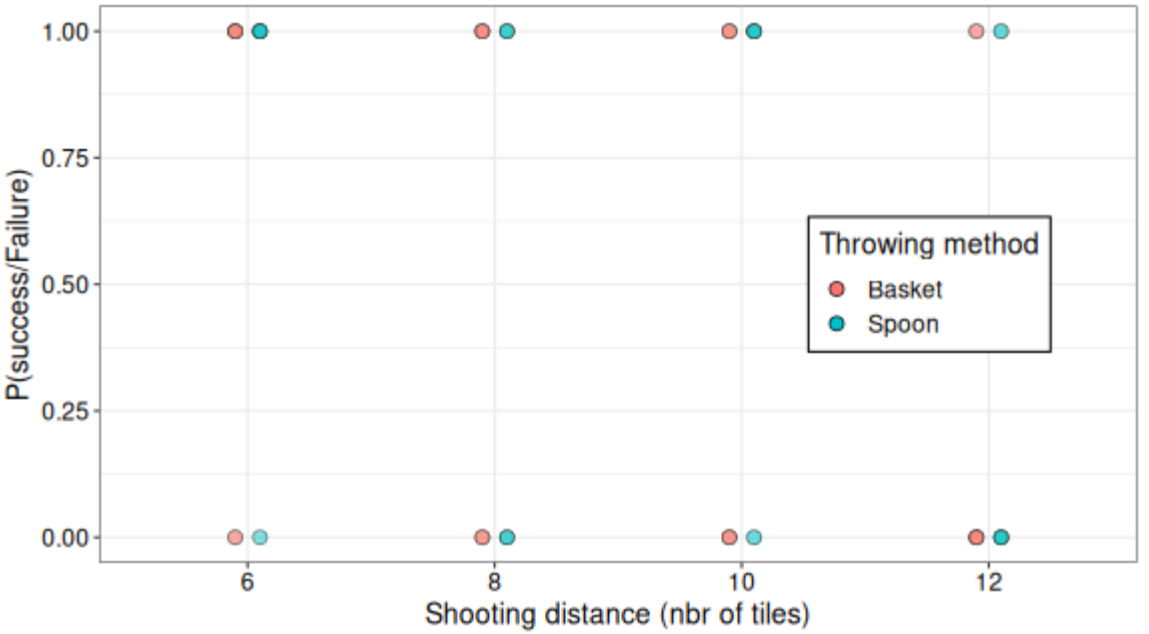


Question: What is the influence of distance (and method) on succes ?



What can we see ?

Question: What is the influence of distance (and method) on succes ?



What can we see ?

- Distance seems to decrease success probability.
- There is no clear effect of the method of throwing.

Reminder: "classical" linear model

$$E(Y) = X\beta + \epsilon$$

β is fitted by the **Ordinary Least Square** (OLS), reduce $\sum \epsilon^2$.

Reminder: "classical" linear model

$$E(Y) = X\beta + \epsilon$$

β is fitted by the **Ordinary Least Square** (OLS), reduce $\sum \epsilon^2$.

X the explanatory variables (predictors):

- Continuous = *regression*
- Discrete = *ANOVA*
- Discrete and continuous = *ANCOVA*

Reminder: "classical" linear model

$$E(Y) = X\beta + \epsilon$$

β is fitted by the **Ordinary Least Square** (OLS), reduce $\sum \epsilon^2$.

X the explanatory variables (predictors):

- Continuous = *regression*
- Discrete = *ANOVA*
- Discrete and continuous = *ANCOVA*

Assumptions:

1. Homogeneity (or homoscedasticity) of variances
2. Normality of **residues**
3. No outliers
4. Data are randomly selected and are independent

"Classical" linear model limits

Does not work in numerous situation. Here are three examples:

"Classical" linear model limits

Does not work in numerous situation. Here are three examples:

Count data:

- Example: plant aboundance
- residual variance increase with count
- $Var(Y)$ increase with $E(Y)$
- **Name:** Poisson

"Classical" linear model limits

Does not work in numerous situation. Here are three examples:

Count data:

- Example: plant aboundance
- residual variance increase with count
- $Var(Y)$ increase with $E(Y)$
- **Name:** Poisson

Fail/Success data:

- Example: presence/absence; win/lose; etc
- $Var(Y)$ local gauss distribution with the two $E(Y)$
- **Name:** Binomial

"Classical" linear model limits

Does not work in numerous situation. Here are three examples:

Count data:

- Example: plant aboundance
- residual variance increase with count
- $Var(Y)$ increase with $E(Y)$
- **Name:** Poisson

Fail/Success data:

- Example: presence/absence; win/lose; etc
- $Var(Y)$ local gauss distribution with the two $E(Y)$
- **Name:** Binomial

Survival data:

- Example: life span expectation
- $Var(Y)$ increase faster than $E(Y)$
- **Name:** Gamma

Math details

GLM general equation:

$$E(Y) = g^{-1}(X\beta) + \epsilon$$

g the link function (e.g. $g(Y) = X\beta$ is the gaussian model)

g allow to calculate the predictions from the linear model.

Math details

GLM general equation:

$$E(Y) = g^{-1}(X\beta) + \epsilon$$

g the link function (e.g. $g(Y) = X\beta$ is the gaussian model)

g allow to calculate the predictions from the linear model.

Distribution	**Example**	**Name**	g	g^{-1}
Normal	Continuous measure	identity	$X\beta = \mu$	$\mu = X\beta$
Poisson	Count	log	$X\beta = \log(\mu)$	$\mu = e^{X\beta}$
Binomial	Nbr: successes/failures	logit	$X\beta = \log(\frac{\mu}{1-\mu})$	$\mu = \frac{e^{X\beta}}{1+e^{X\beta}}$
Gamma	life span expectation	Inverse	$X\beta = \frac{1}{\mu}$	$\mu = \frac{1}{X\beta}$

GLM assumptions

Less assumptions than gaussian linear model:

GLM assumptions

Less assumptions than gaussian linear model:

1. Data are randomly selected and are independent
2. No outliers

GLM assumptions

Less assumptions than gaussian linear model:

1. Data are randomly selected and are independent
2. No outliers

No assumptions on residuals normality:

- Estimate with **Maximum Likelihood** (ML), not OLS
  = Probability to observe Y with the model parameters by iterative process
  = often `log`: log-likelihood or log-like.

GLM assumptions

Less assumptions than gaussian linear model:

1. Data are randomly selected and are independent
2. No outliers

No assumptions on residuals normality:

- Estimate with **Maximum Likelihood** (ML), not OLS
  = Probability to observe Y with the model parameters by iterative process
  = often `log`: log-likelihood or log-like.

No assumptions on homoscedasticity:

- $Var(Y)$ is related to $E(Y)$.
- $Var(Y) = constant \times V(E(Y))$
- V is the function of variance

Additionnal math details: function of variance

Distribution	Writing format	Mean $E(Y)$	Variance $Var(Y)$	Function V
Normal	$N(\mu, \sigma)$	μ	σ	1
Poisson	$P(\lambda)$	λ	λ	λ
Binomial	$B(n, p)$	$E(\frac{Y}{n}) = p$	$Var(\frac{Y}{n}) = \frac{p \times (1-p)}{n}$	$p \times (1-p)$
Gamma	$\Gamma(\mu, k)$	μ	$\frac{\mu^2}{k}$	μ^2

Using the `glm()` function

Using the `glm()` function

In the Basket/Spoon data, it was a Success/Failure experiment: **Binomial data**.

Using the `glm()` function

In the Basket/Spoon data, it was a Success/Failure experiment: **Binomial data**.

```
modg <- glm(ss ~ dist, Baskspo, family = "binomial")
summary(modg)
```

```
##
## Call:
## glm(formula = ss ~ dist, family = "binomial", data = Baskspo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5585  -1.2582   0.8393   1.0211   1.4614
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8622     0.2736   3.151  0.00162 **
## dist8        -0.6742     0.3714  -1.815  0.06946 .
## dist10       -0.4827     0.3737  -1.292  0.19640
## dist12       -1.5089     0.3796  -3.975 7.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 352.64  on 255  degrees of freedom
## Residual deviance: 334.83  on 252  degrees of freedom
## AIC: 342.83
##
```

How estimates are calculated

Use the g function.

How estimates are calculated

Use the g function.

Exemple with the binomial law.

Average success for each distance:

```
Average <- Baskspo %>%
  group_by(dist) %>%
  summarise(Mean = mean(ss))
Average
```

```
## # A tibble: 4 × 2
##   dist   Mean
##   <fct> <dbl>
## 1 6     0.703
## 2 8     0.547
## 3 10    0.594
## 4 12    0.344
```

How estimates are calculated

Use the g function.

Exemple with the binomial law.

Average success for each distance:

```
Average <- Baskspo %>%
  group_by(dist) %>%
  summarise(Mean = mean(ss))
Average
```

```
## # A tibble: 4 × 2
##   dist   Mean
##   <fct> <dbl>
## 1 6     0.703
## 2 8     0.547
## 3 10    0.594
## 4 12    0.344
```

Intercept (Log odd): $\log(p/(1-p))$

```
log(0.703125/(1-0.703125))
```

```
## [1] 0.8622235
```

```
modg$coefficients[1]
```

```
## (Intercept)
##   0.8622235
```

Slope (Log odd-ratio): $\log(\frac{p/(1-p)}{q/(1-q)})$

With: $q = 1 - p$

Model deviance

Log-likelihood difference between our model and the saturated model.

Model deviance

Log-likelihood difference between our model and the saturated model.

In a saturated model, there is one parameter per measurement (Df=0)

Model deviance

Log-likelihood difference between our model and the saturated model.

In a saturated model, there is one parameter per measurement (Df=0)

In all GLM

$$D^2 = 1 - \frac{residual\ deviance}{null\ deviance}$$

- Total model deviance: `modg$null.deviance`

```
## [1] 352.638
```

- Residual deviance: `modg$deviance`

```
## [1] 334.8346
```

D^2 =

```
## [1] 0.05048639
```

Model deviance

Log-likelihood difference between our model and the saturated model.

In a saturated model, there is one parameter per measurement (Df=0)

In all GLM

$$D^2 = 1 - \frac{residual\ deviance}{null\ deviance}$$

- Total model deviance: `modg$null.deviance`

```
## [1] 352.638
```

- Residual deviance: `modg$deviance`

```
## [1] 334.8346
```

D^2 =

```
## [1] 0.05048639
```

In Gaussian LM

OLS R^2 is either:

- **Data variability explained by the model**
- Prediction improvement compared to null model
- Correlation between observed and predicted values

$$R^2 = \frac{SS_{exp}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

Over dispersion

Model deviance follow a χ^2 law:

- $D = \sum_{i=1}^{n} d_i$
- We should observe: $D_{res} = Df_{res}$

Over dispersion

Model deviance follow a χ^2 law:

- $D = \sum_{i=1}^{n} d_i$
- We should observe: $D_{res} = Df_{res}$

If $D_{res} >> Df_{res}$: **Over dispersion**

Solution for over dispersion

Why over dispersion:

- Important predictors not inclued
- Wrong link function (actual Y distribution =/ from real)
- Outliers

Solution for over dispersion

Why over dispersion:

- Important predictors not inclued
- Wrong link function (actual Y distribution =/ from real)
- Outliers

Instead of Maximum Likelihood: **Quasi Maximum Likelihood**

Solution for over dispersion

Why over dispersion:

- Important predictors not inclued
- Wrong link function (actual Y distribution =/ from real)
- Outliers

Instead of Maximum Likelihood: **Quasi Maximum Likelihood**

In `family=` argument: `"quasi..."`

Linear Mixt Effects Models

Dataset: A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios

Run the following script

```
maize_data <- read.table("Data/Maize_data/2a-GrainYield_components_Plot_level.csv"
                         header = T, sep = ",") %>%
  dplyr::select(-type)

maize_data$year <- as.factor(maize_data$year)
maize_data$Replicate <- as.factor(maize_data$Replicate)
maize_data$block <- as.factor(maize_data$block)
maize_data$Row <- as.factor(maize_data$Row)
maize_data$Column <- as.factor(maize_data$Column)
maize_data$Code_ID <- as.factor(maize_data$Code_ID)
maize_data$Variety_ID <- as.factor(maize_data$Variety_ID)
```

```
# colnames(maize_data)
# maize_data$Accession
# glm()
```

Research question:

Previous question (N°.5) we answered with a tow-ways ANOVA: What is the effect of the water treatment and Variety ID `HMV5422`,`11430`, and `F712`, on plant height in 2012?

We had found that the watered plant heights were significantly different to the the rainfed, and that Variety `HMV5422` produced significantly higher plant heights than Variety `11430` in 2012.

Research question:

Previous question (N°.5) we answered with a tow-ways ANOVA: What is the effect of the water treatment and Variety ID `HMV5422`,`11430`, and `F712`, on plant height in 2012?

We had found that the watered plant heights were significantly different to the the rainfed, and that Variety `HMV5422` produced significantly higher plant heights than Variety `11430` in 2012.

Now, let's broaden the research question. What is the effect of the treatment on plant height in the entire experiment?

How do we take into account the fact that the observations came from different sites? years?

Research question:

Previous question (N°.5) we answered with a tow-ways ANOVA: What is the effect of the water treatment and Variety ID `HMV5422`,`11430`, and `F712`, on plant height in 2012?

We had found that the watered plant heights were significantly different to the the rainfed, and that Variety `HMV5422` produced significantly higher plant heights than Variety `11430` in 2012.

Now, let's broaden the research question. What is the effect of the treatment on plant height in the entire experiment?

How do we take into account the fact that the observations came from different sites? years?

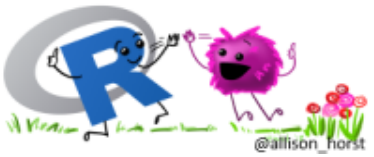
We can use a mixed-effect model using these variables as **random effects**.

Data structure

```
str(maize_data)
```

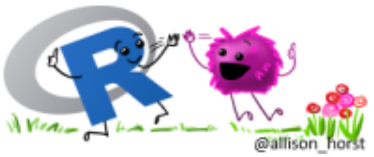
```
## 'data.frame':    19358 obs. of  21 variables:
##  $ Site                          : chr  "Gaillac" "Gaillac" "Gaillac" "Gaillac" ...
##  $ year                          : Factor w/ 3 levels "2011","2012",..: 2 2 2 2 2 2 2 2 2 2 ..
##  $ Experiment                    : chr  "Gai12R" "Gai12R" "Gai12R" "Gai12W" ...
##  $ plotID                        : chr  "G72-01-1-1" "G72-01-2-1" "G72-01-3-1" "G72-02-1-1" ..
##  $ treatment                     : chr  "rainfed" "rainfed" "rainfed" "watered" ...
##  $ Replicate                     : Factor w/ 5 levels "1","2","3","4",..: 1 2 3 1 2 1 2 3 1 2 
##  $ block                         : Factor w/ 69 levels "1","2","3","4",..: 9 7 3 26 18 15 8 22
##  $ Row                           : Factor w/ 138 levels "1","2","3","4",..: 3 11 19 7 14 4 11 
##  $ Column                        : Factor w/ 63 levels "1","2","3","4",..: 5 16 21 11 13 20 27
##  $ Accession                     : chr  "B73_H" "B73_H" "B73_H" "B73_H" ...
##  $ Code_ID                       : Factor w/ 260 levels "3001","3002",..: 1 1 1 1 1 2 2 2 2 2 
##  $ Variety_ID                    : Factor w/ 256 levels "11430","A3","A310",..: 20 20 20 20 20
##  $ plant.height                  : num  170 170 165 255 280 165 165 175 230 230 ...
##  $ tassel.height                 : num  230 220 235 315 340 205 220 215 285 285 ...
##  $ ear.height                    : num  105 105 90 115 160 105 105 80 110 105 ...
##  $ anthesis                      : num  64.8 69.2 67 66 69.2 ...
##  $ silking                       : num  78.9 78.9 81.7 64.8 68.1 ...
##  $ anthesis.silking.interval: num  -14.12 -9.71 -14.63 1.19 1.07 ...
##  $ grain.number                  : num  1492 1521 1960 4462 3987 ...
##  $ grain.yield                   : num  3.41 3.24 3.4 12.55 10.56 ...
##  $ grain.weight                  : num  229 213 174 281 265 ...
```

Let's remove NA values from the plant.height column



Hint: use the piping %>% and the `drop_na()` function.

Let's remove NA values from the plant.height column



Hint: use the piping %>% and the `drop_na()` function.

```
maize_data <- maize_data %>%
  drop_na(plant.height)
```

Fixed vs random effect

Fixed effect:
When the researcher decides the treatments which will be tested.

- ex: effect of three specific varieties on plant variety, to know which variety will produce the tallest plants
- In most studies, the effects are fixed.

Fixed vs random effect

Fixed effect:
When the researcher decides the treatments which will be tested.

- ex: effect of three specific varieties on plant variety, to know which variety will produce the tallest plants
- In most studies, the effects are fixed.

Random effect:
Due design constraints, the researcher randomly selects treatments (e.g. plots, individuals, etc) that will be studied, among all of the available treatments.

- Random effect affect Y variance
- ex: effect of the plant variety on plant height (here, the researcher randomly selects different varieties)
- Here, we are interested in the variability between different varieties, as opposed to specific differences between three Varieties

What's the difference then?

The difference between fixed and random effects is in the interpretation.

- For a fixed effect, the conclusion is applicable to the treatments studied in the experiment (i.e. Varieties `HMV5422`, `11430` and `F712`)
- For a random effect, the conclusion is applicable to all of the Varieties

We also calculate the ANOVA table differently - the denominator used to calculate the F-ratio for the test differs if a factor is random or fixed

Random effects continued

Random effects can also be used to take into account your data structure and statistical independence.

- For example, observations from the same site and the same year are more similar than from a different site or year.

They allow us to take into account pseudoreplication to ensure statistical independence for our observations.

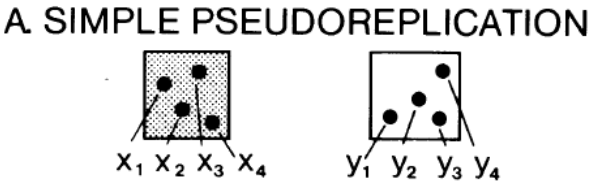
Pseudoreplication

Important paper for experimental set-up: Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. Ecological monographs, 54(2), 187-211.

Three main types of pseudoreplication:

- simple
- sacrificial
- temporal

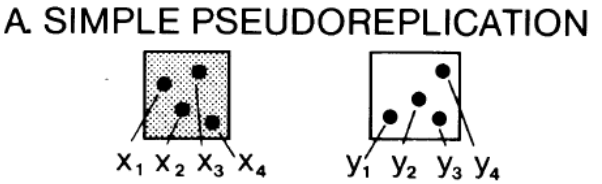
Simple pseudoreplication



Hurlbert et al. 1984, Ecological Monographs

- When there are several observations but only 1 experimental unit per treatment (i.e. if there was only one site with watered vs rainfed, and we took several observations of plant height in each treatment)
- Here, the observations are not independent

Simple pseudoreplication



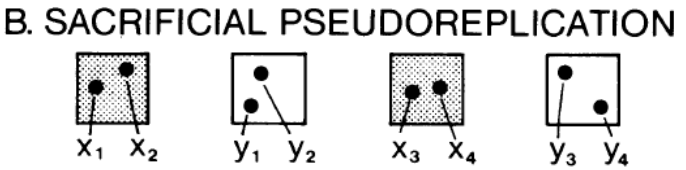
Hurlbert et al. 1984, Ecological Monographs

- When there are several observations but only 1 experimental unit per treatment (i.e. if there was only one site with watered vs rainfed, and we took several observations of plant height in each treatment)
- Here, the observations are not independent

Solution:

- Increase the number of experimental units. This is usually done by having several 'blocks' of each watered and rainfed treatments (in this experiment, there were 2 watered and 3 rainfed experimental units)

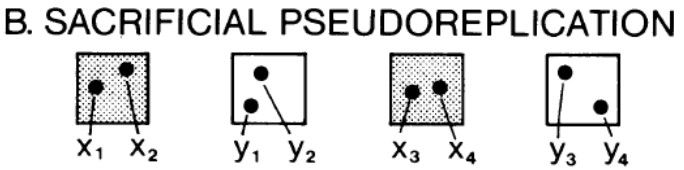
Sacrificial pseudoreplication



Hurlbert et al. 1984, Ecological Monographs

- When there are several observations on an single experimental unit, and/or
- When you combine the data in one analysis without taking into account their origin (i.e. combining data, justifying that "there were no difference between sites so we combined them in one analysis)

Sacrificial pseudoreplication



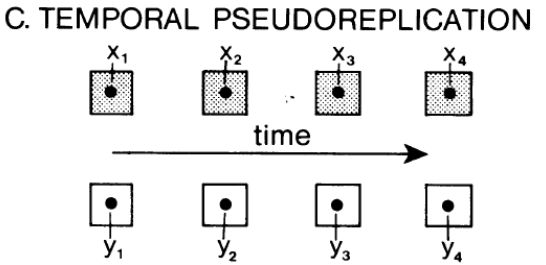
Hurlbert et al. 1984, Ecological Monographs

- When there are several observations on an single experimental unit, and/or
- When you combine the data in one analysis without taking into account their origin (i.e. combining data, justifying that "there were no difference between sites so we combined them in one analysis)

Solution:

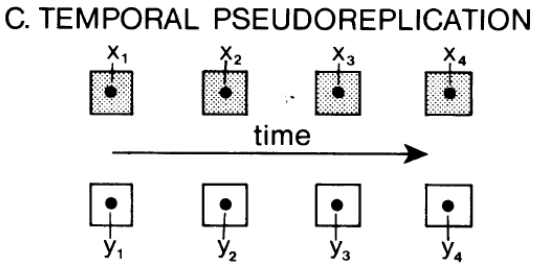
- Use a mean of the observations on one experimental unit in the model, or
- Use a model which takes into account the strucutre that there are observations within an experimental unit (i.e. using random effects)

Temporal pseudoreplication



- When you take several measurements on a single experimental unit (i.e. measuring the plant height once a month), and that you consider these measurements to be independent

Temporal pseudoreplication

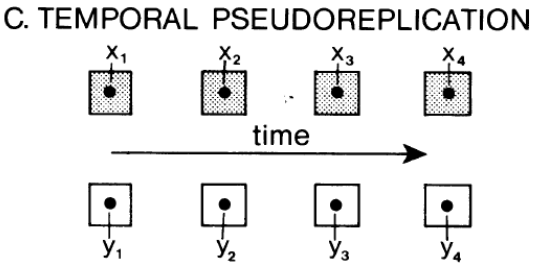


- When you take several measurements on a single experimental unit (i.e. measuring the plant height once a month), and that you consider these measurements to be independent

Solution:

- Analyse each the measurements from each month (or other time period) separately
- Use a model which takes into account repeated measures (ANOVA for repeated measures, mixed models, etc. to take into account the autocorrelation between different measurements).

Temporal pseudoreplication



- When you take several measurements on a single experimental unit (i.e. measuring the plant height once a month), and that you consider these measurements to be independent

Solution:

- Analyse each the measurements from each month (or other time period) separately
- Use a model which takes into account repeated measures (ANOVA for repeated measures, mixed models, etc. to take into account the autocorrelation between different measurements).

We will not go over these, but there plenty of resources online.

Linear Mixed Effect Models (LMM)

Mixed effect models include a large variety of different models which allows us to correctly take into account the data structure (nested data/ grouping factors, repeated measurements, etc.)

Linear Mixed Effect Models (LMM)

Mixed effect models include a large variety of different models which allows us to correctly take into account the data structure (nested data/ grouping factors, repeated measurements, etc.)

Let's go back to our question for this week: is there an effect of the treatment on plant height in our experiment (several sites, 2 years and 37 varieties)?

Why use a linear mixed effects model (LMM)?

Linear Mixed Effect Models (LMM)

Mixed effect models include a large variety of different models which allows us to correctly take into account the data structure (nested data/ grouping factors, repeated measurements, etc.)

Let's go back to our question for this week: is there an effect of the treatment on plant height in our experiment (several sites, 2 years and 37 varieties)?

Why use a linear mixed effects model (LMM)?

- Here, we are interested in the general effect
- The relationship may differ slightly among varieties due to unmeasured processes, or among experimental sites or year due to unmeasured environmental variables. We want to represent this data structure in our model.

Why choose a LMM?

LMM are a balance between separating the dataset (per site, year, etc) and lumping the data together (i.e. not accounting for differences between the sites, etc)

- Estimate slope and intercept parameters for each site and year (separating) but estimate fewer parameters than a classical regression.
- Use all the data available (lumping) while accounting for pseudoreplication and controlling for differences among sites and years.

How do LMMs work?

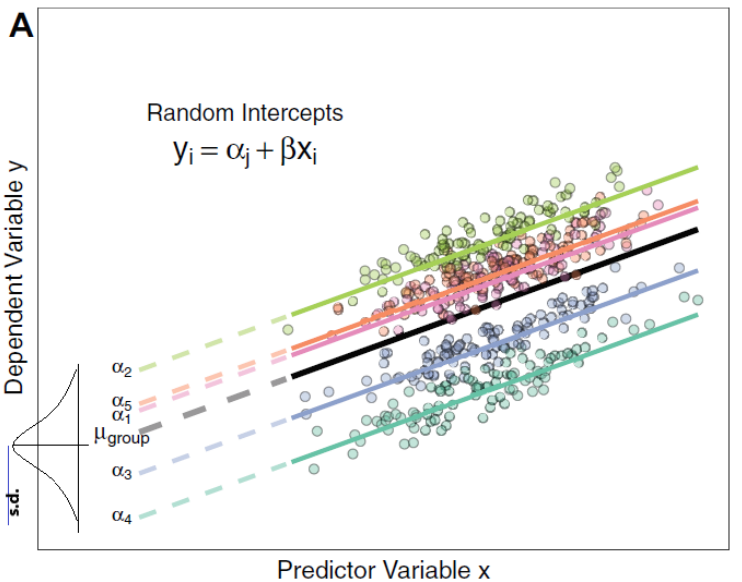
- Intercepts and/or slopes are allowed to vary according to a given factor (i.e. random effect factor), such as site or year
- Intercepts, slopes and their confidence interval are adjusted to take into account the data structure

How do LMMs work?

- Intercepts and/or slopes are allowed to vary according to a given factor (i.e. random effect factor), such as site or year
- Intercepts, slopes and their confidence interval are adjusted to take into account the data structure
- LMM are gaussian models
- GLMM also exist, you can 'combine' GLM and LMM

Random intercept

- It is assumed that the intercepts come from a normal distribution
- Only need to estimate the mean (μ) and standard deviation of the normal distribution instead of the n intercepts (i.e. one for each site)

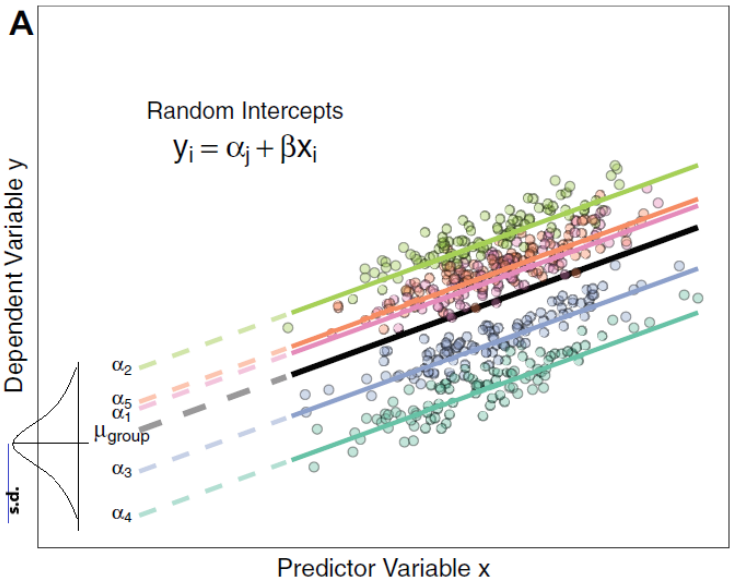


Harrison et al. 2018, PeerJ

Note that the more levels your factor has, the more accurately the mean and standard deviation of the normal distribution will be estimated.

Random intercept

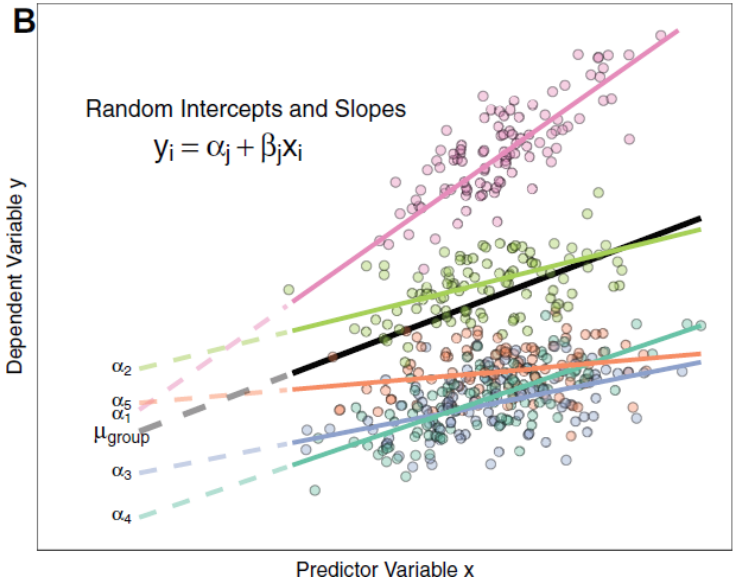
Thus, the model only needs to estimate the mean and standard distribution of the intercepts, instead of the 29 intercepts (for sites)



Harrison et al. 2018, PeerJ

Random slope

The same principle applies to slopes that vary according to a given factor (i.e. the random effect of site differs on the rainfed vs watered treatments) - only the mean and s.d. of the slopes are estimated.

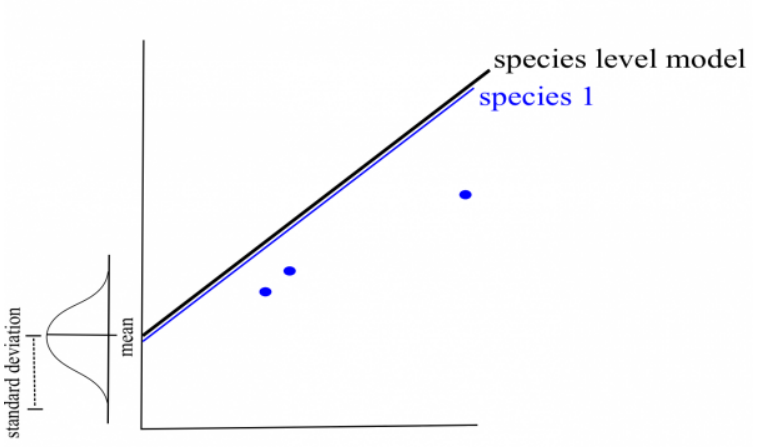


Harrison et al. 2018, PeerJ

Here, both intercepts and slopes are permitted to vary by group. Random slope models give the model far more flexibility to fit the data, but require a lot more data to obtain accurate estimates of separate slopes for each group.

Taking into account the data structure

- If a certain site or year is poorly represented (not many values), the model will give more weight to the pooled model to estimate the intercept and slope of that site or year.



@CSBQ

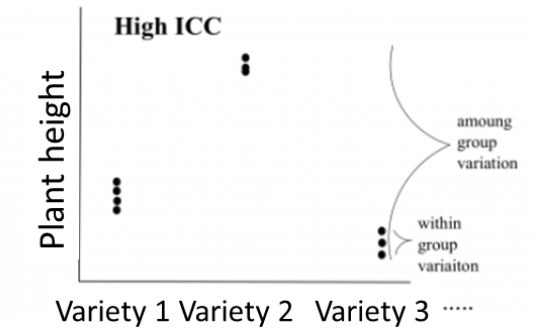
Taking into account the data structure

The confidence intervals for the intercepts and slopes are adjusted to take account of the pseudo-replication-based on the **intraclass correlation coefficient (ICC)**

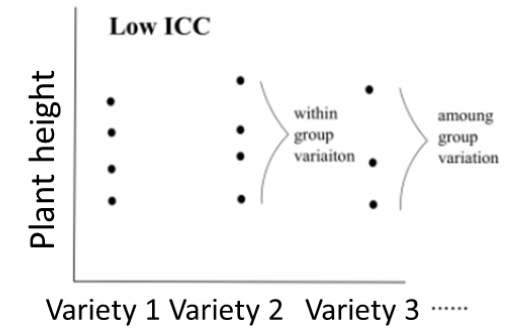
ICC: How much variation is there in each group versus between groups?

Interclass correlation coefficient (ICC)

High ICC

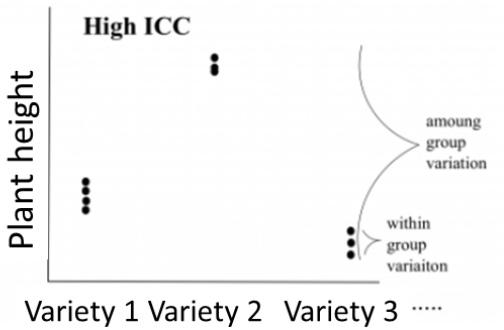


Low ICC

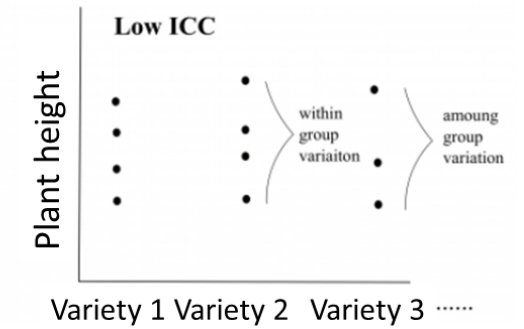


Interclass correlation coefficient (ICC)

High ICC



Low ICC

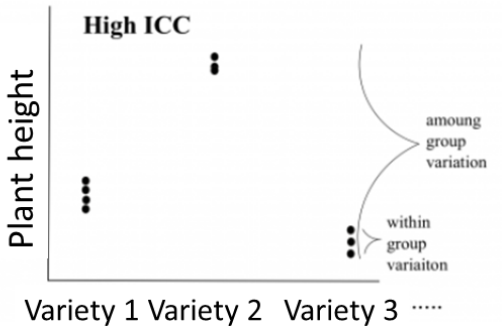


High ICC (low variation within group, and high variation among groups)

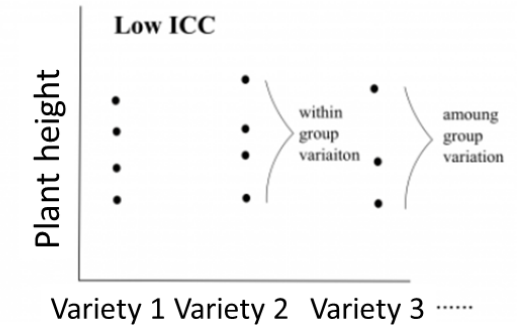
- points are treated as single observation because they are correlated
- small effective sample size
- **large confidence intervals** for slope and intercept

Interclass correlation coefficient (ICC)

High ICC



Low ICC



High ICC (low variation within group, and high variation among groups)

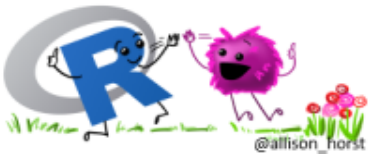
- points are treated as single observation because they are correlated
- small effective sample size
- **large confidence intervals** for slope and intercept

Low ICC

- points coming from the same Variety are treated independently because they are little correlated
- large effective sample size
- **small confidence intervals** for slope and intercept

Data exploration

Look at the distribution of samples for each factor level using the `table()` function or the `replications()` function:



Data exploration

Look at the distribution of samples for each factor level using the `table()` function or the `replications()` function:


```
table(maize_data$Site)
```

```
##
##     Bologna  Campagnola     Craiova    Debrecen     Gaillac    Graneros
##        1194        2698        1253        2189        2518        1264
##   Karlsruhe Martonvasar      Murony       Nerac
##        2701        1260        1260        2991
```

```
table(maize_data$year)
```

```
##
## 2011 2012 2013
## 1075 8730 9523
```

```
table(maize_data$treatment)
```

```
##
## rainfed watered
##   11930    7398
```

Data exploration

Mixed-effect models can be used to analyze unbalanced experimental plans

Data exploration

Mixed-effect models can be used to analyze unbalanced experimental plans

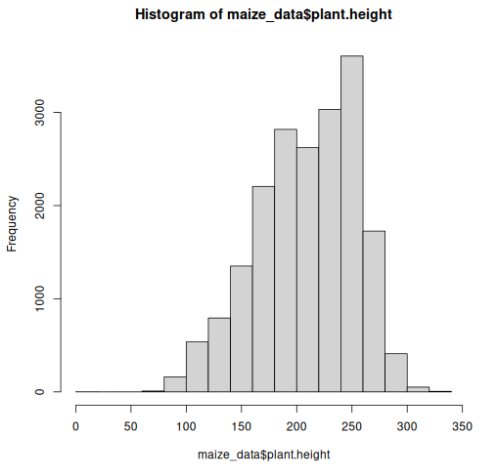
Look at the distribution of the continuous variables using the `hist()` function

Data exploration

Mixed-effect models can be used to analyze unbalanced experimental plans

Look at the distribution of the continuous variables using the `hist()` function

```
hist(maize_data$plant.height)
```



Major deviations could cause heteroscedasticity problems. If necessary, make transformations. In this case, the data seems OK.

Data exploration

Check for collinearity between your explanatory variables

Ex: If we wanted to test the effect of plant height and tassel height on grain weight.

Data exploration

Check for collinearity between your explanatory variables

Ex: If we wanted to test the effect of plant height and tassel height on grain weight.

- The problem with collinear predictors is simply that they explain the same thing, so their effect on the response variable will be confounded in the model
- In this example, there is no risk of collinearity with no continuous variables. If you had another continuous variable (Var2), one simple way to check for collinearity is:

```
cor(var1, var2)
```

In the above example, it would be better to just include plant **or** tassel height on grain weight

Setting up the model

Let's go back to our original question: what is the effect of treatment on the plant height?

Which are our fixed effect factors? And our random factors?

Setting up the model

Let's go back to our original question: what is the effect of treatment on the plant height?

Which are our fixed effect factors? And our random factors?

Fixed effects

- `treatment`: this is something we controlled and specifically want to test

Random effects

- `Variety_ID`: here, we are interested in the **general** trend of Variety, not in specific differences between the chosen varieties.
- `Site`
- `year`
- `Replicate` within the `treatment` / `year` / `Site`

Depending on the research question, random effects could be fixed effects

How to write an LMM in R?

There are several packages which can be used: `lme4` or `nlme`. Today we will look at the `lmer()` (linear mixed model) function from the `lme4`package.

```
mod_lmer <- lmer(plant.height ~ treatment +
                   (1|Variety_ID) + (1|Site) + (1|year) +
                   (1|Variety_ID:Site) +
                   (1|Variety_ID:year) +
                   (1|Variety_ID:treatment),
                 data = maize_data, REML = TRUE)
```

```
## boundary (singular) fit: see help('isSingular')
```

- `(1|Variety_ID)`: indicates varying intercept but keeping the same slope
- `:` : indicates an interaction effect
- `REML = TRUE`: estimation method

Note: Here we have **not** added the random effect (1|treatment/year/Site/Replicate) because the model fails to converge (too many parameters to estimate).

Note on estimation methods

REML (Restricted Maximum Likelihood) is the default method in `lmer`.

Note on estimation methods

REML (Restricted Maximum Likelihood) is the default method in `lmer`.

Note that the standard deviation estimator in the Maximum Likelihood (ML) is biased by a factor of (n−2)/n, especially on small dataset. The REML method corrects this bias.

Note on estimation methods

REML (Restricted Maximum Likelihood) is the default method in `lmer`.

Note that the standard deviation estimator in the Maximum Likelihood (ML) is biased by a factor of (n−2)/n, especially on small dataset. The REML method corrects this bias.

- We should compare nested random effect models with REML (such as `treatment/year/Site/Replicate`)

  Note: REML takes into accounts the number of estimated parameters and loses 1 Df per parameter

- While we should compare nested fixed effects models with ML

  Note: work only if the fixed effects are the same

What if we wanted the slope of a random effect to vary?

Let's say that we think that the random effect of the site will be dependent on the water treatment (i.e. a site in southern Europe may influence plant.height differently when rainfed than a northern site which receives more rain).

```
mod_lmer2 <- lmer(plant.height ~ treatment +
                    (1|Variety_ID) + (treatment|Site) +
                     (1|year) +
                     (1|Variety_ID:Site) +
                     (1|Variety_ID:year) +
                     (1|Variety_ID:treatment),
                  data = maize_data, REML = TRUE)
```

We will continue with the previous model.

A note on model selection

The choice of the factors which are included in the model depends on the research question.

However, to determine if you have built the best mixed model based on your prior knowledge, you should compare this *a priori* model to other alternative models

With the dataset we are working on, there are several alternative models that might better fit the data.

Model selection

We can see if our model compares to the basic linear model which does not include random factors. To do so, we need to change the estimate method to ML, so `REML = FALSE` because `lm()` doesn't use the same estimation method as `lmer()`.

For example, we could compare the following models (we will skip this step):

```
#Linear model with no random effects
M0 <- lm(plant.height ~ treatment, data = maize_data)

#Our model
M1 <- lmer(plant.height ~ treatment + (1|Variety_ID) +
             (1|Site) + (1|year) + (1|Variety_ID:Site) +
             (1|Variety_ID:year) + (1|Variety_ID:treatment),
           data = maize_data, REML = FALSE)

#Lmer model with Experiment and Replicate
M2 <- lmer(plant.height ~ treatment + (1 | Variety_ID) +
             (1 | Experiment/Replicate), data = maize_data, REML = FALSE)

#Lmer model with varying intercepts and slopes
M3 <- lmer(plant.height ~ treatment + (treatment | Variety_ID) +
             (treatment | Experiment/Replicate), data = maize_data, REML = FALSE)
```

Model selection

- Models can be compared by using the AICc function from the `AICcmodavg` package
- The Akaike Information Criterion (AIC) is a measure of model quality that can be used to compare models
- AICc corrects for bias created by small sample sizes

More information for model selection can be found here: https://qcbsrworkshops.github.io/workshop06/workshop06-en/workshop06-en.html#57

--

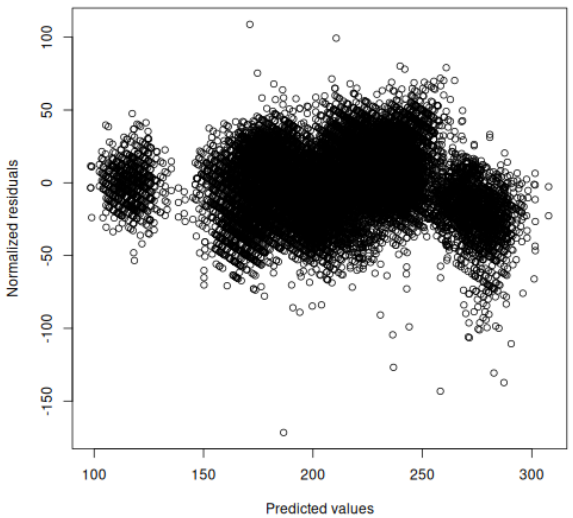
We will skip this process today, and continue with our original `mod_lmer` model.

Check the model assumptions

1. Homogeneity of variance (predicted values vs residual values plot)
2. Check independence of the model residuals
3. Check normality of model residuals (but mixed-models are robust to deviations from normality)

Homogeneity of variance

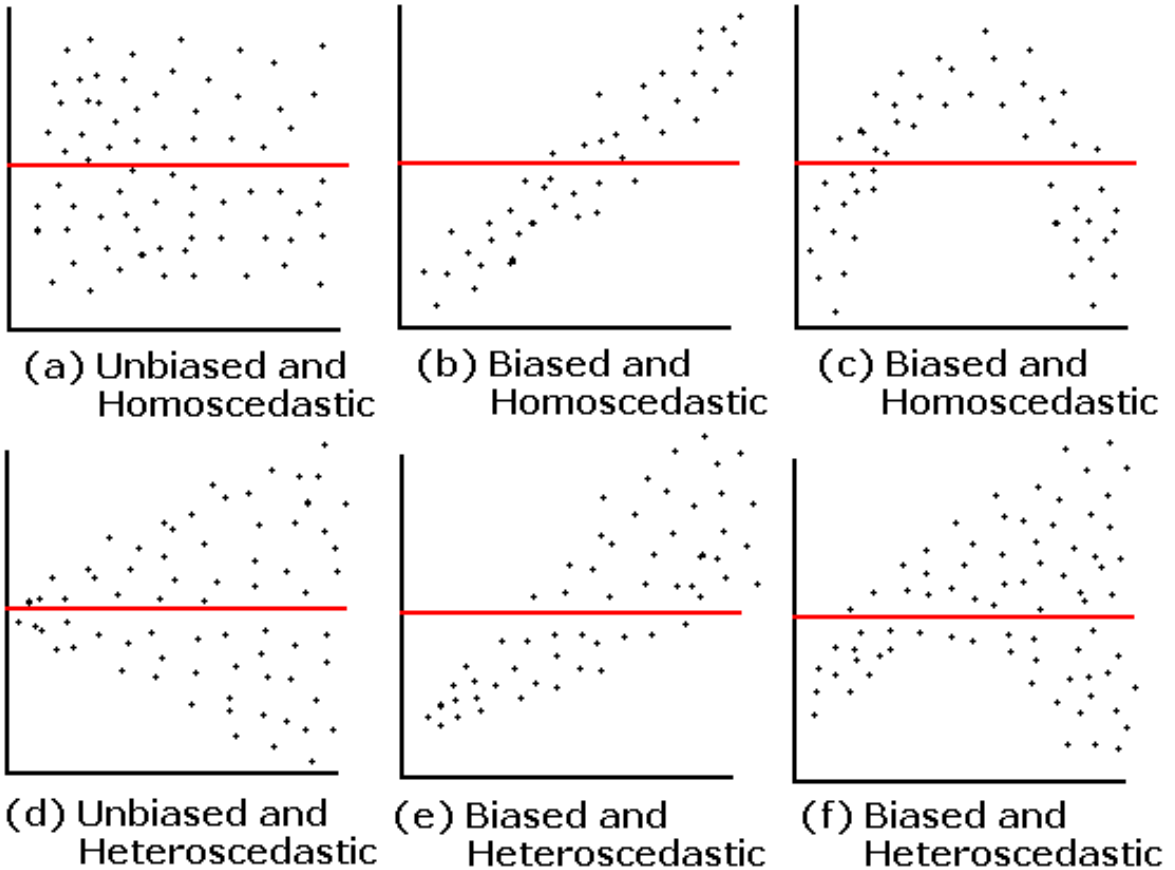
```
plot(resid(mod_lmer) ~ fitted(mod_lmer),
     xlab = 'Predicted values', ylab = 'Normalized residuals')
```



There seems to be a few outliers, but no large trends. We will keep all data points.

However, if we wanted to remove some points we could use the function `identify(resid(mod_lmer)~ fitted(mod_lmer))` and click on the points, which gives us the row number of the individuals from the data table

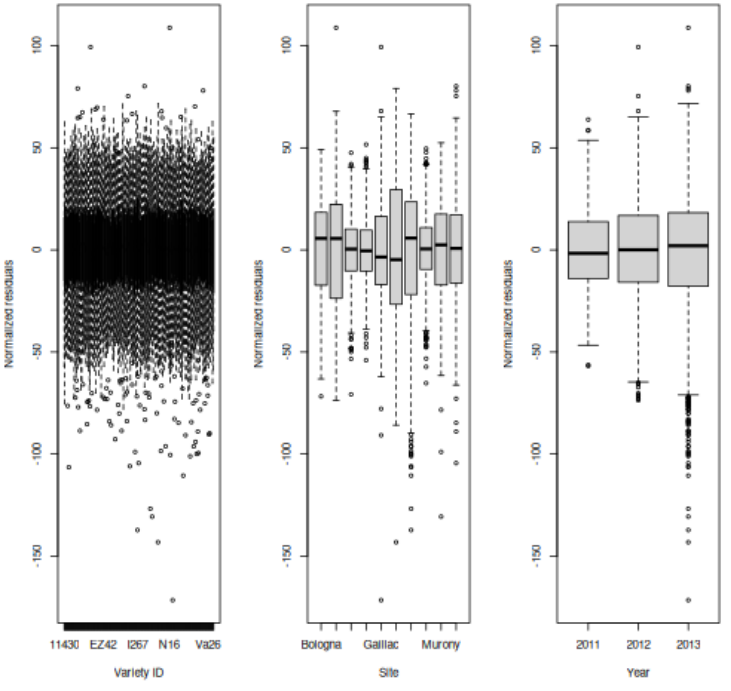
Homogeneity of variance



@CSBQ

Independence of model residuals with each covariate

```
par(mfrow = c(1,3)) #to get a good window to see the graphs
boxplot(resid(mod_lmer) ~ Variety_ID, data = maize_data,
        xlab = "Variety ID", ylab = "Normalized residuals")
boxplot(resid(mod_lmer) ~ Site, data = maize_data,
        xlab = "Site", ylab = "Normalized residuals")
boxplot(resid(mod_lmer) ~ year, data = maize_data,
        xlab = "Year", ylab = "Normalized residuals")
```



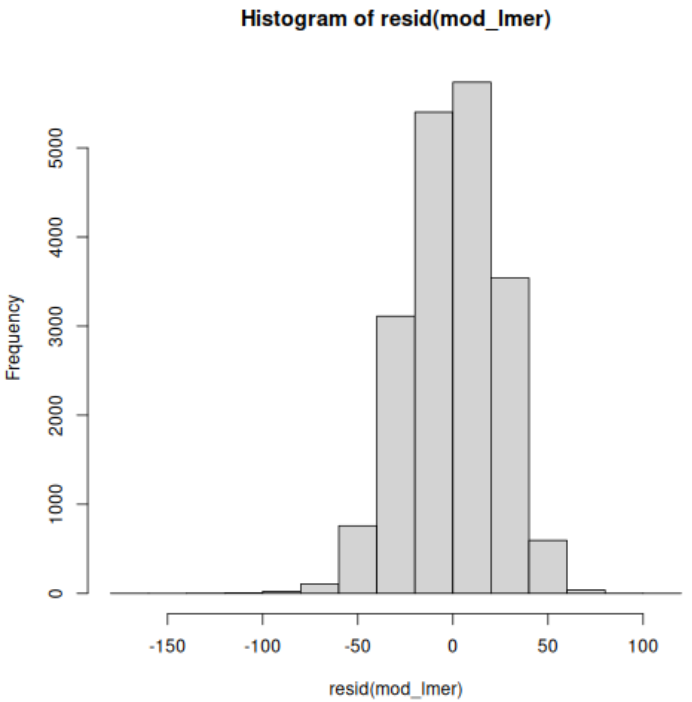
[see boxplots on previous slide]

Here, we want to check for a homogeneous dispersion of the residuals around 0, i.e. that there is no pattern of residuals depending on the variable

The assumption is respected here, but we could consider removing some data points from 2013 or from certain sites. We will continue with our current data set. (The code for the figures will be available online)

Normality of model residuals

```
hist(resid(mod_lmer))
```



Residuals follow a normal distribution, which indicates that the model is not biased and over-influenced by certain values.

Model interpretation

```
summary(mod_lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: plant.height ~ treatment + (1 | Variety_ID) + (1 | Site) + (1 |
##     year) + (1 | Variety_ID:Site) + (1 | Variety_ID:year) + (1 |
##     Variety_ID:treatment)
##    Data: maize_data
##
## REML criterion at convergence: 178226.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.1645 -0.6883  0.0332  0.7309  4.5451
##
## Random effects:
##  Groups               Name        Variance  Std.Dev.
##  Variety_ID:Site      (Intercept) 0.000e+00 0.000e+00
##  Variety_ID:year      (Intercept) 0.000e+00 0.000e+00
##  Variety_ID:treatment (Intercept) 6.153e-08 2.481e-04
##  Variety_ID           (Intercept) 5.196e+01 7.208e+00
##  Site                 (Intercept) 1.058e+03 3.253e+01
##  year                 (Intercept) 7.430e+01 8.620e+00
##  Residual                         5.736e+02 2.395e+01
## Number of obs: 19328, groups:
## Variety_ID:Site, 2537; Variety_ID:year, 597; Variety_ID:treatment, 512; Variety_ID, 256;
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      197.1733    11.4424   17.23
```

Fixed effects

You can select the fixed effects from the general model (all explicative variables and interactions) with the `dredge()` function from the `MuMIn` package.

Fixed effects

You can select the fixed effects from the general model (all explicative variables and interactions) with the `dredge()` function from the `MuMIn` package.

In `lme4` package, the author purposefully did not include p-values.

The `nlme` package provide p-values.

Fixed effects

You can select the fixed effects from the general model (all explicative variables and interactions) with the `dredge()` function from the `MuMIn` package.

In `lme4` package, the author purposefully did not include p-values.

The `nlme` package provide p-values.

There is a discussion on how to calculate *Df* and therefore p-values in LMM.

You can use the `Anova()` function from the `car` package, but it ignore the problem with *Df*

Fixed effects

You can select the fixed effects from the general model (all explicative variables and interactions) with the `dredge()` function from the `MuMIn` package.

In `lme4` package, the author purposefully did not include p-values.

The `nlme` package provide p-values.

There is a discussion on how to calculate *Df* and therefore p-values in LMM.

You can use the `Anova()` function from the `car` package, but it ignore the problem with *Df*

The way we interpret the results is by looking at the estimated slope of the fixed effect +/- the 95% confidence interval (if we set our alpha = 0.05).

```
Fixed effects:
                 Estimate Std. Error t value
(Intercept)      197.1733    11.4450   17.23
treatmentwatered  42.7005     0.3564  119.82
```

Fixed effects: confidence intervals

The `confint()` function calculates the confidence intervals for the fixed effect, and for the sigmas, which correspond to the random effects.

```
confint(mod_lmer)
```

Note: this will take a long time to run.

```
> confint(mod_lmer)
Computing profile confidence intervals ...
                    2.5 %      97.5 %
.sig01           0.000000   0.6071411
.sig02           0.000000   0.7858577
.sig03           0.000000   1.4289936
.sig04           6.534170   7.9777834
.sig05          21.229947  52.0799507
.sig06           4.029294  27.2346340
.sigma          23.710361  24.1911409
(Intercept)    175.345889 219.0080324
treatmentwatered  42.002083  43.3990452
```

```
Fixed effects:
                 Estimate Std. Error t value
(Intercept)      197.1733    11.4450   17.23
treatmentwatered  42.7005     0.3564  119.82
```

If the 95% confidence interval of the slope does not include 0, the slope (here 42.7, seen as the `Estimate` in the fixed effects table), and therefore the effect of the treatment, is signficantly different from 0 at the threshold alpha = 0.05.

Mixed effect models

- Report the fixed effect estimates and the confidence limits: "The effect of the watering treatment on plant height is strong and confidence intervals are narrow"
- Report how variable the effect is between different random effects: "On average the effect is strong, but there is considerably variation betwen sites, much more than between Variety types"

```
Random effects:
 Groups               Name        Variance Std.Dev.
 Variety_ID:Site      (Intercept)    0.00   0.000
 Variety_ID:year      (Intercept)    0.00   0.000
 Variety_ID:treatment (Intercept)    0.00   0.000
 Variety_ID           (Intercept)   51.97   7.209
 Site                 (Intercept) 1058.93  32.541
 year                 (Intercept)   74.33   8.621
 Residual                          573.57  23.949
Number of obs: 19328, groups:
Variety_ID:Site, 2537; Variety_ID:year, 597; Variety_ID:treatment, 512; Variety_ID,
 256; Site, 10; year, 3
```

Mixed effect models

- Report the fixed effect estimates and the confidence limits: "The effect of the watering treatment on plant height is strong and confidence intervals are narrow"
- Report how variable the effect is between different random effects: "On average the effect is strong, but there is considerably variation betwen sites, much more than between Variety types"

```
Random effects:
 Groups               Name        Variance Std.Dev.
 Variety_ID:Site      (Intercept)    0.00   0.000
 Variety_ID:year      (Intercept)    0.00   0.000
 Variety_ID:treatment (Intercept)    0.00   0.000
 Variety_ID           (Intercept)   51.97   7.209
 Site                 (Intercept) 1058.93  32.541
 year                 (Intercept)   74.33   8.621
 Residual                          573.57  23.949
Number of obs: 19328, groups:
Variety_ID:Site, 2537; Variety_ID:year, 597; Variety_ID:treatment, 512; Variety_ID,
 256; Site, 10; year, 3
```

Note: the `Correlation of Fixed Effects` in the summary is the correlation between estimated coefficients for each fix effect.

Heritability

Heritability

For many plant breeding applications, we consider the main effects to be random (such as `Variety_ID`), and want to estimate the proportion of variance due to these effects on a certain variable (i.e. `plant.height`) in our experimental design.

We can use this information to calculate heritability.

Heritability

For many plant breeding applications, we consider the main effects to be random (such as `Variety_ID`), and want to estimate the proportion of variance due to these effects on a certain variable (i.e. `plant.height`) in our experimental design.

We can use this information to calculate heritability.

Definition: the heritability is the proportion of phenotypic variability explained by genetic variability.

The **broad-sense heritability** (H^2) can be calculated with a mixed-model allowing us to estimate the V_G (genetic variance) and the V_E (environmental variance):

$$H^2 = \frac{V_G}{V_P} = \frac{V_G}{(V_G+V_E)/nrep}$$

Heritability

For many plant breeding applications, we consider the main effects to be random (such as `Variety_ID`), and want to estimate the proportion of variance due to these effects on a certain variable (i.e. `plant.height`) in our experimental design.

We can use this information to calculate heritability.

Definition: the heritability is the proportion of phenotypic variability explained by genetic variability.

The **broad-sense heritability** (H^2) can be calculated with a mixed-model allowing us to estimate the V_G (genetic variance) and the V_E (environmental variance):

$$H^2 = \frac{V_G}{V_P} = \frac{V_G}{(V_G+V_E)/nrep}$$

- V_P is the phenotypic variability: $V_P = V_G + V_E + Cov(G, E)$.
  Note: when genotypes are not related to specific environment, $Cov(G, E) = 0$.
- `nrep` being the mean number of repetition for one genotype in the experiment.
- H^2 comprise between 0 and 1, with 0 no variability due to genetic.

Heritability

For many plant breeding applications, we consider the main effects to be random (such as `Variety_ID`), and want to estimate the proportion of variance due to these effects on a certain variable (i.e. `plant.height`) in our experimental design.

We can use this information to calculate heritability.

Definition: the heritability is the proportion of phenotypic variability explained by genetic variability.

The **broad-sense heritability** (H^2) can be calculated with a mixed-model allowing us to estimate the V_G (genetic variance) and the V_E (environmental variance):

$$H^2 = \frac{V_G}{V_P} = \frac{V_G}{(V_G+V_E)/nrep}$$

- V_P is the phenotypic variability: $V_P = V_G + V_E + Cov(G, E)$.
  Note: when genotypes are not related to specific environment, $Cov(G, E) = 0$.
- `nrep` being the mean number of repetition for one genotype in the experiment.
- H^2 comprise between 0 and 1, with 0 no variability due to genetic.

We can extract this information from the summary table.

Let's go back to our model:

```
summary(mod_lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: plant.height ~ treatment + (1 | Variety_ID) + (1 | Site) + (1 |
##     year) + (1 | Variety_ID:Site) + (1 | Variety_ID:year) + (1 |
##     Variety_ID:treatment)
##    Data: maize_data
##
## REML criterion at convergence: 178226.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.1645 -0.6883  0.0332  0.7309  4.5451
##
## Random effects:
##  Groups               Name        Variance  Std.Dev.
##  Variety_ID:Site      (Intercept) 0.000e+00 0.000e+00
##  Variety_ID:year      (Intercept) 0.000e+00 0.000e+00
##  Variety_ID:treatment (Intercept) 6.153e-08 2.481e-04
##  Variety_ID           (Intercept) 5.196e+01 7.208e+00
##  Site                 (Intercept) 1.058e+03 3.253e+01
##  year                 (Intercept) 7.430e+01 8.620e+00
##  Residual                         5.736e+02 2.395e+01
## Number of obs: 19328, groups:
## Variety_ID:Site, 2537; Variety_ID:year, 597; Variety_ID:treatment, 512; Variety_ID, 256;
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)      197.1733    11.4424   17.23
## treatmentwatered  42.7005     0.3564  119.82
##
## Correlation of Fixed Effects:
```

Extracting the variance

Extract the Variance using the following function:

```
print(VarCorr(mod_lmer), comp="Variance")
```

```
##  Groups                Name        Variance
##  Variety_ID:Site       (Intercept) 0.0000e+00
##  Variety_ID:year       (Intercept) 0.0000e+00
##  Variety_ID:treatment (Intercept) 6.1529e-08
##  Variety_ID            (Intercept) 5.1961e+01
##  Site                  (Intercept) 1.0584e+03
##  year                  (Intercept) 7.4304e+01
##  Residual                          5.7357e+02
```

Extracting the variance

Extract the Variance using the following function:

```
print(VarCorr(mod_lmer), comp="Variance")
```

```
##  Groups                Name        Variance
##  Variety_ID:Site       (Intercept) 0.0000e+00
##  Variety_ID:year       (Intercept) 0.0000e+00
##  Variety_ID:treatment  (Intercept) 6.1529e-08
##  Variety_ID            (Intercept) 5.1961e+01
##  Site                  (Intercept) 1.0584e+03
##  year                  (Intercept) 7.4304e+01
##  Residual                          5.7357e+02
```

Store the Variance, using the following function:

```
sigmas <- as.data.frame( VarCorr( mod_lmer) )$vcov
```

Now we have stored our information in a list:

```
print(VarCorr(mod_lmer), comp="Variance")
```

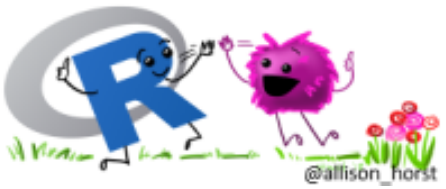
```
##  Groups                Name        Variance
##  Variety_ID:Site       (Intercept) 0.0000e+00
##  Variety_ID:year       (Intercept) 0.0000e+00
##  Variety_ID:treatment  (Intercept) 6.1529e-08
##  Variety_ID            (Intercept) 5.1961e+01
##  Site                  (Intercept) 1.0584e+03
##  year                  (Intercept) 7.4304e+01
##  Residual                          5.7357e+02
```

```
print(sigmas)
```

```
## [1] 0.000000e+00 0.000000e+00 6.152918e-08 5.196111e+01 1.058429e+03
## [6] 7.430429e+01 5.735723e+02
```

Our `sigmas` list is in the same order as the `VarCorr()` of our model.

Calculating H^2



$$H^2 = \frac{V_G}{(V_G+V_E)/nrep}$$

Calculate the H^2, given that V_G is the variance of the `Variety_ID`, and V_E is the sum of the variance of the environmental effects in interaction (`Variety_ID:Site`, `Variety_ID:year`, `Variety_ID:treatment`) and of the residuals

- In this estimation of heritability, we are ignoring the main random effects

Hint:

- use the `[]` to choose the number in the `sigmas` data frame, by looking at the order in the `print(VarCorr(mod_lmer), comp="Variance")` table
- use classical operators (/ , + , and the`sum()` function)
- to find the nrep of each group, look at the `summary(mod_lmer)` table

Solution

```
H2 <- sigmas[4] / #VG = Variety_ID in the 4th position of the list
  sum( sigmas[4], #VG
       sigmas[1]/10, #because 10 sites
       sigmas[3]/2, #because 2 treatments
       sigmas[2]/2, #because 2 years
       sigmas[7]/(2*2*10)) #residual divided by number of sites*treatments*years
H2
```

```
## [1] 0.7837222
```

The 0.78 value is a relatively high H^2 : 78% of phenotypic variability is due to genetic variability. It could be possible since the study likely used pre-selected Varieties which have a high yield. This indicates that plant height is a highly heritable trait.

--

Note: there are several ways to code the model and calculate heritability.That's where it can get quite complicated! For example, if we calculated an H^2 using the main effects instead of the interaction effects, we would have an H^2 of 0.23.

References

Mazerolle, M. J. *VII - Blocs*. FOR7044 Analyse de Données. Université Laval, Automne 2019.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. Ecological monographs, 54(2), 187-211.

LMMs:

- https://wiki.qcbs.ca/r_workshop6
- https://campus.datacamp.com/courses/hierarchical-and-mixed-effects-models-in-r/linear-mixed-effect-models?ex=7
- Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., ... & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ, 6, e4794
- Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M. (2011). Mixed effects models and extensions in ecology with R, Statistics for biology and health. Springer, New York, NY.

Heritability:

- https://dyerlab.github.io/Landscape-Genetics-Data-Analysis/quantitative-genetics.html#heritability
- https://www.youtube.com/watch?v=LqhNkwVcH-Q&t=411s

Next session:

- Data visualization
- Making reproducible graphics with the `ggplot2` package