Statistical testing

Session 2 and 3 LOVE'R course

Pablo Raguet

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

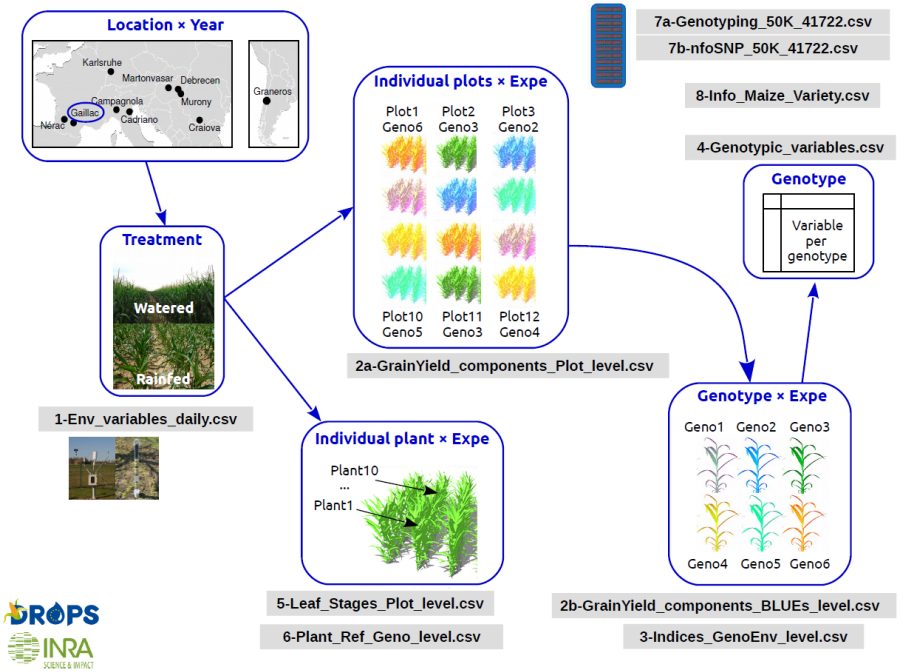
original teacher: Tania L. Maxwell
website: tania-maxwell.github.io

10-11-2022

Topics for today

- Explain the dataset we will be working with
- Data format
- Visual data
- Treatment effects
- What to extract from your analysis

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



Citation: Millet, Emilie J.; Pommier, Cyril; Buy, Mélanie; Nagel, Axel; Kruijer, Willem; Welz-Bolduan, Therese; Lopez, Jeremy; Richard, Cécile; Racz, Ferenc; Tanzi, Franco; Spitkot, Tamas; Canè, Maria-Angela; Negro, Sandra S.; Coupel-Ledru, Aude; Nicolas, Stéphane D.; Palaffre, Carine; Bauland, Cyril; Praud, Sébastien; Ranc, Nicolas; Presterl, Thomas; Bedo, Zoltan; Tuberosa, Roberto; Usadel, Björn; Charcosset, Alain; van Eeuwijk, Fred A.; Draye, Xavier; Tardieu, François; Welcker, Claude, 2019, "A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios", https://doi.org/10.15454/IASSTN, Portail Data INRAE, V2, UNF:6:zF9w0A2f+MHeW7maeeXJWA== [fileUNF]

Experiment detail

A panel of 256 maize hybrids was grown:

- with two water regimes (irrigated or rainfed),
- in seven sites spread along a climatic transect from western to eastern Europe
- in 2012 and 2013

which equals 29 experiments, as the combination of one year, one site and one water regime, with two and three repetitions for waterred and rainfed treatments, respectively.

Measurements:

- Hourly records of micrometeorological data and soil water status, and associated with precise measurement of phenology
- **Grain yield and its components at the end of the experiment**

Statistical Analysis

What is the **first** step to do when you start your statistical analysis?

Statistical Analysis

What is the **first** step to do when you start your statistical analysis?

State the specific research question(s)

Your statistical test should be based on the biological question.

Statistical Analysis

What is the **first** step to do when you start your statistical analysis?

State the specific research question(s)

Your statistical test should be based on the biological question.

1. Is there a relation between the grain weight and the tassel height in the Gaillac site?
2. What is the effect of the plant height on the grain yield?
3. What is the effect of the plant height and tassel height on the grain number?
4. What is the effect of the water treatment on plant height?
5. What is the effect of the water treatment and the varity ID on grain weight
6. What are the effect of the grazing treatment and roots length on fruit production?

Today's objectives

- Go through 6 questions together, then individually
- Look into detail our analyses
- Meanwhile, go through important concepts to keep in mind

Data Import

```
library(PerformanceAnalytics)
library(tidyverse)
library(corrplot)
library(multcomp)
library(nlme)
library(lme4)

# not needed in a R project :
# dir <- getwd()
# setwd(dir)


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

Note: Because we are importing a `.csv` file, the `sep =` will be a comma.

Data Import

```
library(PerformanceAnalytics)
library(tidyverse)
library(corrplot)
library(multcomp)
library(nlme)
library(lme4)

# not needed in a R project :
# dir <- getwd()
# setwd(dir)


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

Note: Because we are importing a `.csv` file, the `sep =` will be a comma.

What kind of data do we have here?

Hint: Click on the data in your Global Environment, and use the `str()` function,

Converting our data into the proper format

You could write out each individual column in the following way:

```
maize_data$year <- as.factor(maize_data$year)
```

```
maize_data$Replicate <- as.factor(maize_data$Replicate)
```

etc...

Converting our data into the proper format

You could write out each individual column in the following way:

```
maize_data$year <- as.factor(maize_data$year)

maize_data$Replicate <- as.factor(maize_data$Replicate)
```

etc...

OR you could use `mutate_if()` from the `dplyr` package.

Here, the function states "If a column is a character, convert it as a factor" for the entire data frame.

```
maize_data <- maize_data %>%
  mutate_if(is.character, as.factor)
```

- It is important to this at the beginning to reduce future problems with analyses.
- **Tip:** try to name your columns in a consistent manner (i.e. all variables with capital first letter), because R is case-sensitive

Exploring the data

How many observations do we have per experiment?

Hint: use the `table()` function

Exploring the data

How many observations do we have per experiment?

Hint: use the `table()` function

```
table(maize_data$Experiment)
```

```
##
## Bol12R Bol12W Cam11R Cam11W Cam12R Cam12W Cam13R Cam13W Cra12R Cra12W Deb11R
##    756    449    108     72    756    504    756    504    753    500    105
## Deb11W Deb12R Deb12W Deb13R Gai12R Gai12W Gai13R Gai13W Gra13R Gra13W Kar11R
##     70    756    504    756    756    504    756    504    759    506    148
## Kar11W Kar12R Kar12W Kar13R Kar13W Mar13R Mar13W Mur13R Mur13W Ner11R Ner11W
##    100    747    504    723    487    756    504    756    504    285    190
## Ner12R Ner12W Ner13R Ner13W
##    756    504    756    504
```

What do we see here? Does this make sense? Think of the experiment set up.

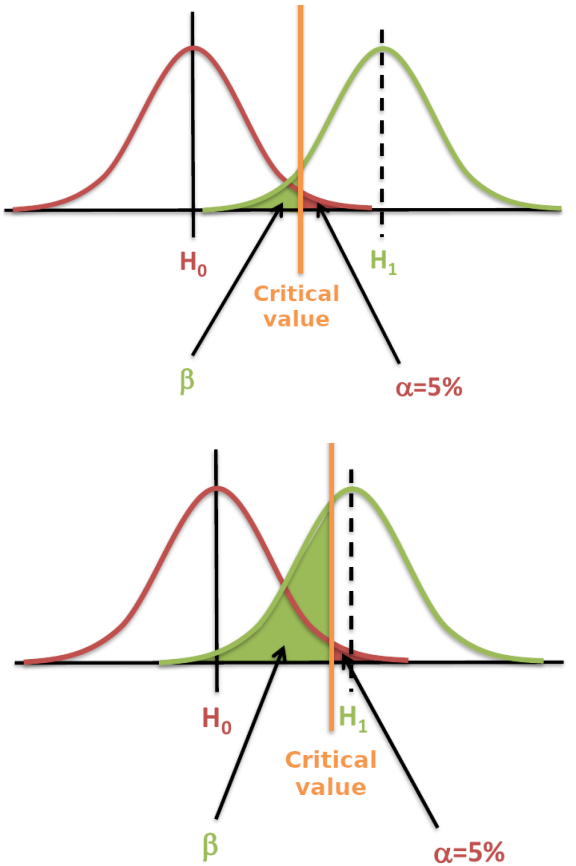
Reminder: some statistical tests

N sample	Var type	Statistical question	Parametric	Non-parametric
1	Qualitative	Adjustment to a theoretical law	Test adjustment (n>5)	Fisher test (n≤5)
	Quantitative	Normality	Shapiro test	
2	Qualitative	Ditributions independence	χ2 independence (n>5)	Fisher test (n≤5)
	Quantitative	Homoscedasticity	Bartlett test	
		Mean comparison	Student test	- Mann-Whitney test (paired samples) - Wilcoxon test (non-paired samples)
		Relation	- Pearson correlation - linear regression	Spearman correlation
>2	Quantitative	Mean comparison	- ANOVA - Tukey test	Kruskal-Wallis test

Reminder: α risk, β risk and power

H_0: null hypothesis, there is no effect.

- **Risk** α (type 1 error):
  P(reject H_0 | H_0 TRUE)
- **Risk** β (type 2 error):
  P(conserve H_0 | H_0 FALSE)
- **Power:**
  P(reject H_0 | H_0 FALSE)
  β = 1 - Power



Question 1: Is there a relation between the grain weight and the tassel height in the Gaillac site?

Question 1: Is there a relation between the grain weight and the tassel height in the Gaillac site?

Key parts to this question:

- Relation between two **numerical** variables
- What are the units?

`grain.weight`: individual grain weight (mg)

`tassel.height`: plant height including tassel, from ground level to the highest point of the tassel (cm)

Thus, they are both **continuous** variables

Reminder


CONTINUOUS
measured data, can have ∞ values within possible range.
I AM 3.1" TALL
I WEIGH 34.16 grams
DISCRETE
observations can only exist at limited values, often counts.
I HAVE 8 LEGS and 4 SPOTS!
@allison_horst

Data Frame

First, separate the data frame to only include values from the Gaillac site.

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

Data Frame

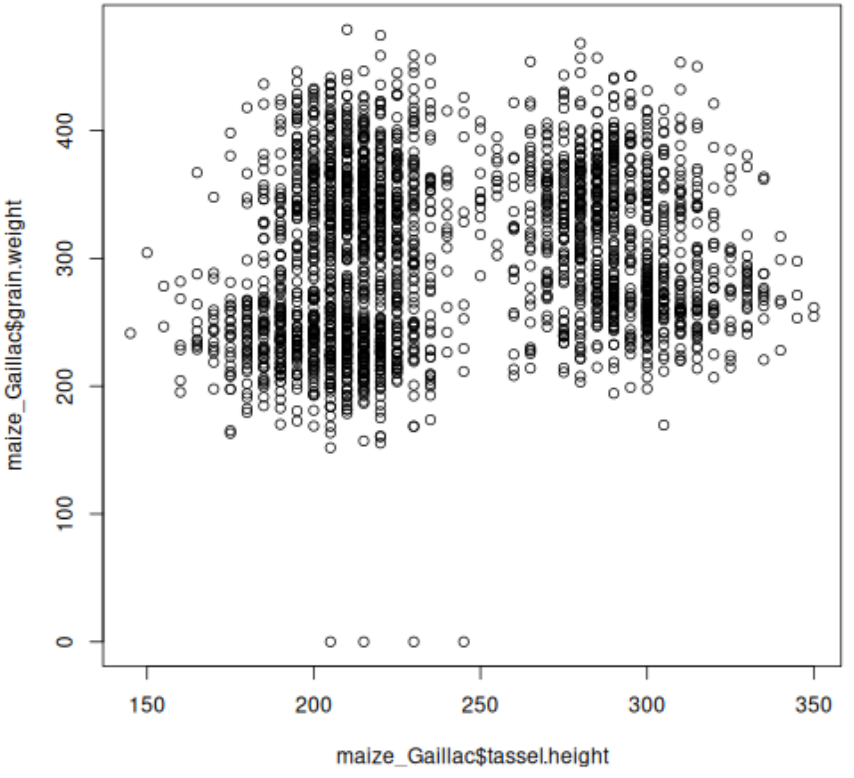
First, separate the data frame to only include values from the Gaillac site.

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

```
maize_Gaillac <- maize_data %>%
  filter(Site == "Gaillac")
```

Visualize the relation between grain weight and tassel height

```
plot(maize_Gaillac$grain.weight ~ maize_Gaillac$tassel.height)
```



Do the statistical test

Which test do you think we should use? Why?

Do the statistical test

Which test do you think we should use? Why?

Pearson's correlation

Why?

- Looking for the **relation** between two variables
- There is not a clear cause and effect relationship between the variables (linear regression)
- There may be a third non-measured variable which acts on both the variables in the correlation

Correlation vs linear regression

Correlation

- used to associate two variables (no cause and effect)
- mostly used as a "pre-analysis" phase to determine relations between explanatory variables (note: in different)

Linear regression

- often used for observational studies
- a cause and effect relationship is difficult to assure without having controled for other pertinente variables
- be careful with interpretation

What are the test assumptions for Pearson's correlation?

1. each variable should be continuous

2. related pairs: each value should have both a grain yield and a tassel height value

3. absence of outliers: can skew

4. linearity: the shape of the dots is a line and not curved

Tip: you can find test assumptions online

Do we satisfy the test assumptions?

[x] **1.** Both variables are continuous

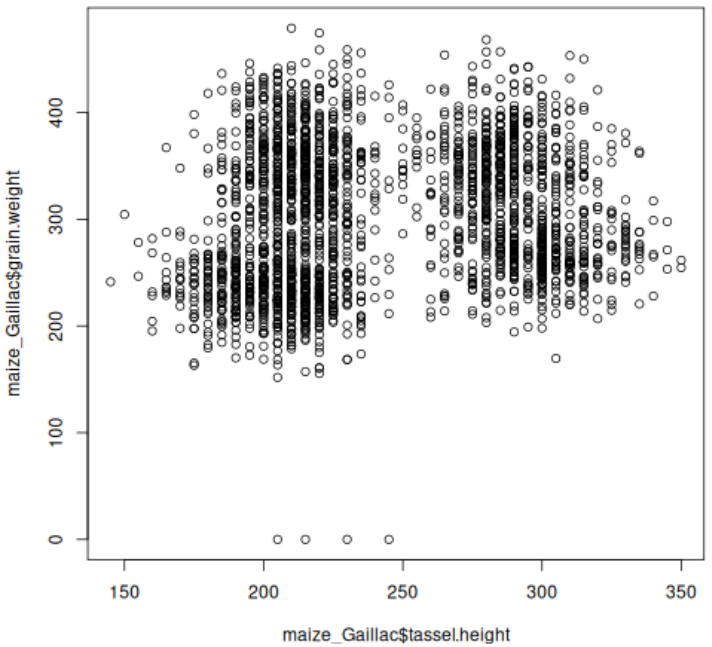
Related pairs? Let's remove the na values from the data frame using the `drop_na()`.

```
maize_Gaillac <- maize_Gaillac %>%
  drop_na("grain.weight", "tassel.height")
```

[x] **2.** NA values removed

Are there outliers which could skew the correlation?

```
plot(maize_Gaillac$grain.weight ~ maize_Gaillac$tassel.height)
```



Visually, we can determine to remove the values at the bottom of the graph, with grain weights lower than 100mg.

Remove the values with a grain weigh lower than 100 mg to create a new data frame for the analyses.

Hint: use the `filter()` function.

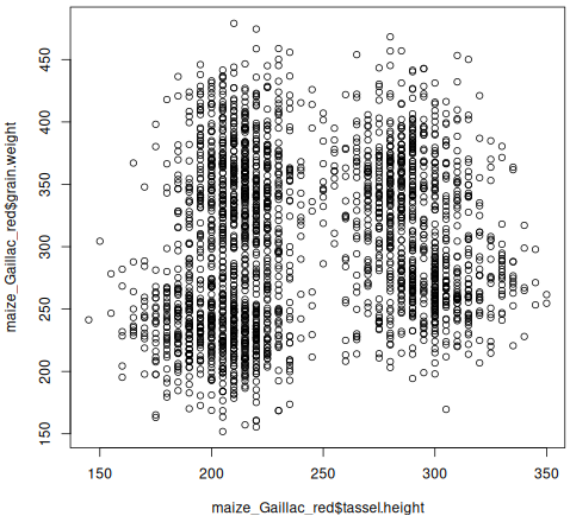
Remove the values with a grain weigh lower than 100 mg to create a new data frame for the analyses.

Hint: use the `filter()` function.

```
maize_Gaillac_red <- maize_Gaillac %>%
  filter(grain.weight>100)
```

Check the number of observations in each data frame in your Global Environment. You can also check with a graph:

```
plot(maize_Gaillac_red$grain.weight ~ maize_Gaillac_red$tassel.height)
```



[x] **3.** Outliers removed

[x] **4.** Linearity (no curved shape within the dots)

Pearson's correlation

Here the p-value corresponds to the statistical test with

$\mathbf{H_0}$: r = 0 (there is no correlation)

$\mathbf{H_1}$: r ≠ 0

Pearson's correlation

Here the p-value corresponds to the statistical test with

$\mathbf{H_0}$: r = 0 (there is no correlation)

$\mathbf{H_1}$: r ≠ 0

```
cor.test(maize_Gaillac_red$grain.weight, maize_Gaillac_red$tassel.height)
```

```
##
##     Pearson's product-moment correlation
##
## data:  maize_Gaillac_red$grain.weight and maize_Gaillac_red$tassel.height
## t = 6.8529, df = 2471, p-value = 9.103e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0976787 0.1750430
## sample estimates:
##       cor
## 0.1365691
```

The Pearson's correltation shows a significant positive (`cor` value is positive) relationship.

However, a significant result is not very informative (with a large number of samples, anything can become significant)

It is better to look at the strength of the correlation (r^2)

However, a significant result is not very informative (with a large number of samples, anything can become significant)

It is better to look at the strength of the correlation (r^2)

Finding the r^2

The r^2 is the r correlation value, squared (i.e. coefficient of determination)

```
cor(maize_Gaillac_red$grain.weight, maize_Gaillac_red$tassel.height)^2
```

```
## [1] 0.01865111
```

We can conclude that **1.87%** of the variation of the grain weight is associated to the tassel height.

Subsetting for a correlation matrix

We can get a quick overview of how all of our numeric columns value relate to one another

First, we need to subset only the numeric columns. One way we can do this is by viewing our data table and noting the numeric columns.

```
maize_Gaillac_num <- maize_Gaillac_red[,c(13:21)]
```

Or, we could select columns which are numeric using `select(where(is.numeric))`

```
maize_Gaillac_num <- maize_Gaillac_red %>%
  dplyr::select(where((is.numeric))) %>%
  dplyr::select(-c(year, Replicate, block, Row, Column, Code_ID))
# The last line remove irrelevant columns
```

Note: the `select()` function is present in several packages, so by using `dplyr::` we are telling R in which package to take the function.

Subsetting for a correlation matrix

We can get a quick overview of how all of our numeric columns value relate to one another

First, we need to subset only the numeric columns. One way we can do this is by viewing our data table and noting the numeric columns.

```
maize_Gaillac_num <- maize_Gaillac_red[,c(13:21)]
```

Or, we could select columns which are numeric using `select(where(is.numeric))`

```
maize_Gaillac_num <- maize_Gaillac_red %>%
  dplyr::select(where((is.numeric))) %>%
  dplyr::select(-c(year, Replicate, block, Row, Column, Code_ID))
# The last line remove irrelevant columns
```

Note: the `select()` function is present in several packages, so by using `dplyr::` we are telling R in which package to take the function.

We need to remove NA values from the data frame: use the `na.omit()` function.

```
maize_Gaillac_num <- na.omit(maize_Gaillac_num)
```

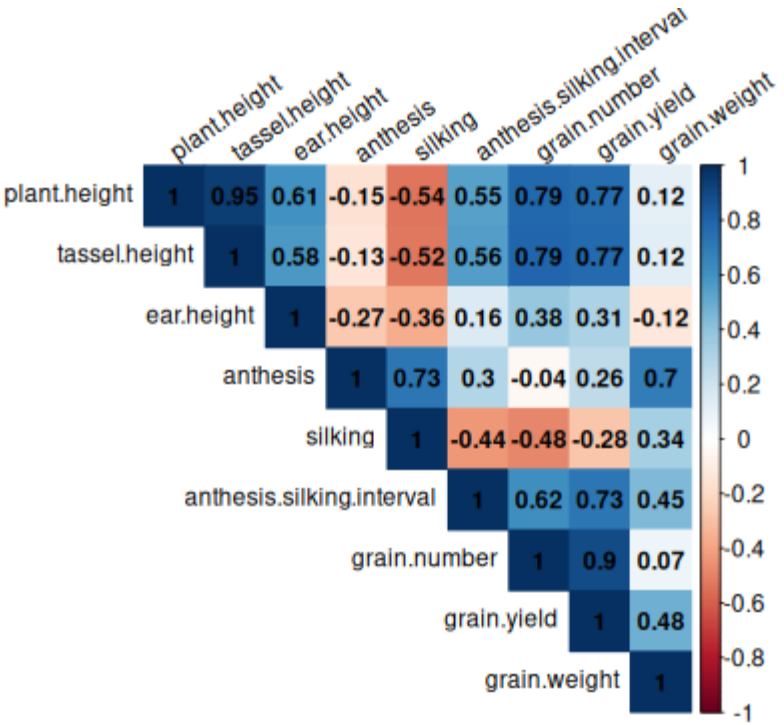
We will use a function which needs a correlation matrix from our values. To do so, we will use the `cor` function.

```
##                            plant.height tassel.height ear.height   anthesis
## plant.height                  1.0000000     0.9497685  0.6078690 -0.1500663
## tassel.height                 0.9497685     1.0000000  0.5773443 -0.1332881
## ear.height                    0.6078690     0.5773443  1.0000000 -0.2670345
## anthesis                     -0.1500663    -0.1332881 -0.2670345  1.0000000
## silking                      -0.5357567    -0.5246874 -0.3649882  0.7285945
## anthesis.silking.interval     0.5492235     0.5558489  0.1576083  0.2989595
##                               silking anthesis.silking.interval grain.number
## plant.height                -0.5357567                 0.5492235   0.78812204
## tassel.height               -0.5246874                 0.5558489   0.79251303
## ear.height                  -0.3649882                 0.1576083   0.38103365
## anthesis                     0.7285945                 0.2989595  -0.03801357
## silking                      1.0000000                -0.4357996  -0.48014072
## anthesis.silking.interval   -0.4357996                 1.0000000   0.61898079
##                            grain.yield grain.weight
## plant.height                 0.7669227    0.1196133
## tassel.height                0.7716080    0.1242338
## ear.height                   0.3138608   -0.1214250
## anthesis                     0.2574101    0.6968856
## silking                     -0.2798214    0.3352556
## anthesis.silking.interval    0.7280936    0.4486556
```

Correlation matrix

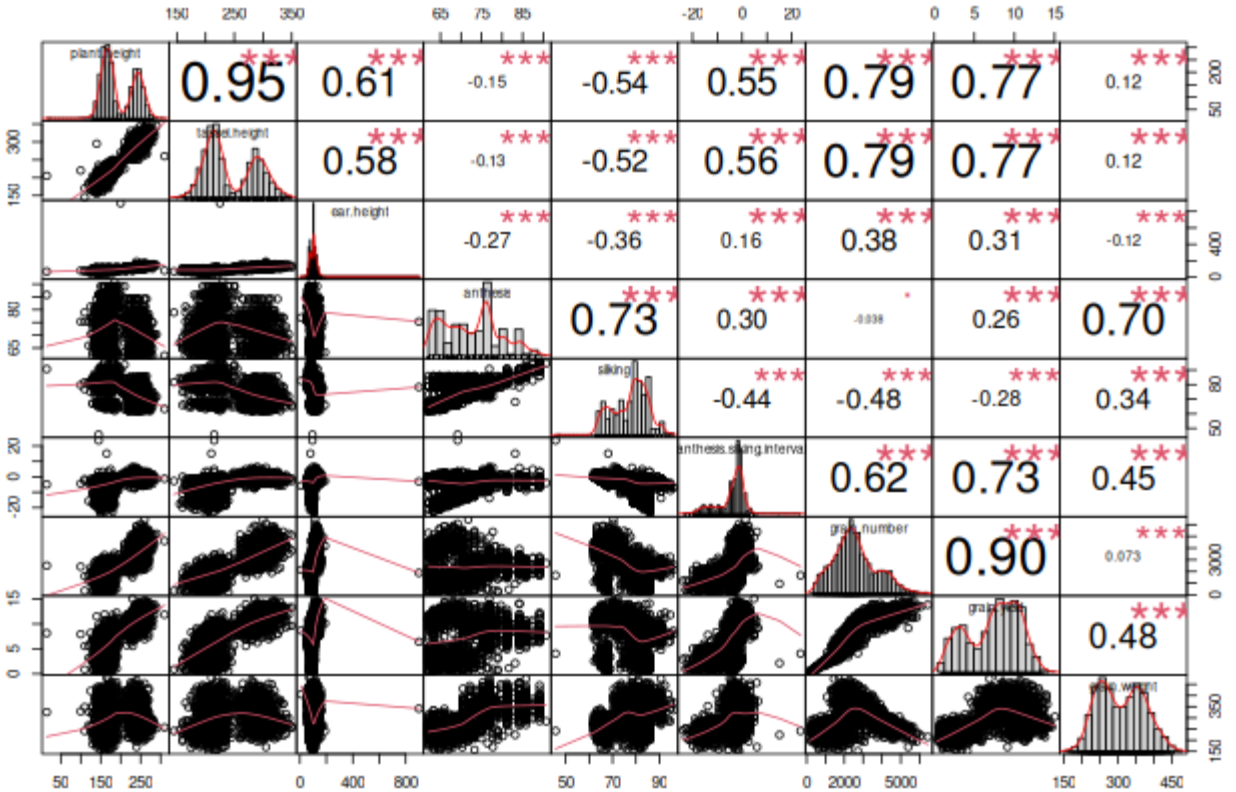
Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients.

```
corrplot(corr_maize, method = "color", type = "upper", addCoef.col = "black",
         tl.srt = 35, tl.col= "black",   cl.cex = 1)
```



Another function

`chart.Correlation()` from the package `PerformanceAnalytics`:



Your turn

Using the previous data table, what is the relation between grain weight and tassel height within the waterred site? And within the rainfed site?

Solution

```
maize_Gaillac_wat <- maize_Gaillac_red %>%
  filter(treatment == "watered")

cor(maize_Gaillac_wat$grain.weight, maize_Gaillac_wat$tassel.height)^2
```

```
## [1] 0.0748907
```

We can conclude that **7.49%** of the variation of the grain weight is associated to the tassel height if we only include the watered sites.

Solution

```
maize_Gaillac_wat <- maize_Gaillac_red %>%
  filter(treatment == "watered")

cor(maize_Gaillac_wat$grain.weight, maize_Gaillac_wat$tassel.height)^2
```

```
## [1] 0.0748907
```

We can conclude that **7.49%** of the variation of the grain weight is associated to the tassel height if we only include the watered sites.

```
maize_Gaillac_rain <- maize_Gaillac_red %>%
  filter(treatment == "rainfed")

cor(maize_Gaillac_rain$grain.weight, maize_Gaillac_rain$tassel.height)^2
```

```
## [1] 0.04363096
```

We can conclude that **4.36%** of the variation of the grain weight is associated to the tassel height if we only include the rainfed sites.

Does this make sense?

For te next analysis: reduce the `maize_data` dataframe

Separate the data frame.

- Only the year 2012
- To simplify our question, we will focus on three Varieties: `"HMV5422"`,`"11430"` and `"F712"`

  Hint: use the `|` symbol to include several values for `Variety_ID`

- Remove the NA values from all the columns

Answer

```
maize_2012 <- maize_data %>%
  filter(year == "2012") %>%
  filter(Variety_ID == "HMV5422" | Variety_ID == "11430" | Variety_ID == "F712")%>%
  drop_na(.) %>%
  droplevels() #we need to add this to remove the other levels of Variety
```

Question 2: What is the effect of the plant height on the grain yield?

Question 2: What is the effect of the plant height on the grain yield?

Key parts to this question:

- Effect of **one numerical** variable on another **numerical** variable
- What are the units?

`plant.height`: plant height from ground to the base of the flag leaf (cm)

`grain.yield`: grain yield adjusted at 15% humidity (t ha^{-1})

They are both **continuous** variables

Question 2: What is the effect of the plant height on the grain yield?

Key parts to this question:

- Effect of **one numerical** variable on another **numerical** variable
- What are the units?

`plant.height`: plant height from ground to the base of the flag leaf (cm)

`grain.yield`: grain yield adjusted at 15% humidity (t ha^{-1})

They are both **continuous** variables

Question: What is the statistical approach we should use ?

What are the Linear Model assumptions?

Before analyzing the data, check the following assumptions:

1. Homogeneity (or homoscedasticity) of variances (look at residuals vs fitted)
2. Normality of **residues**
   (but normality isn't as important as homoscedasticity)
3. No outliers
4. Data are randomly selected and are independent

What are the Linear Model assumptions?

Before analyzing the data, check the following assumptions:

1. Homogeneity (or homoscedasticity) of variances (look at residuals vs fitted)
2. Normality of **residues**
   (but normality isn't as important as homoscedasticity)
3. No outliers
4. Data are randomly selected and are independent

Note: Balanced data should be preferred

Linear model: general equation

Math equation:

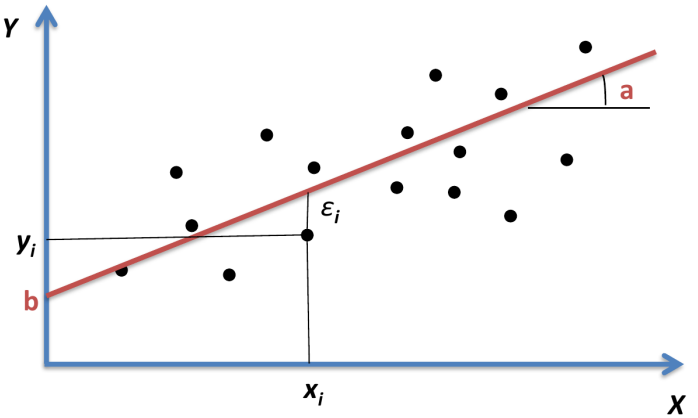
$\hat{Y} = X\beta$

OR

$y_i = a \times x_i + b + \epsilon_i$

With:

- Y and y_i the explained variable
- X and x_i the explanatory variable(s)
- β the model parameters :
  - a the slope
  - b the intercept
- ϵ_i the model residuals (error)



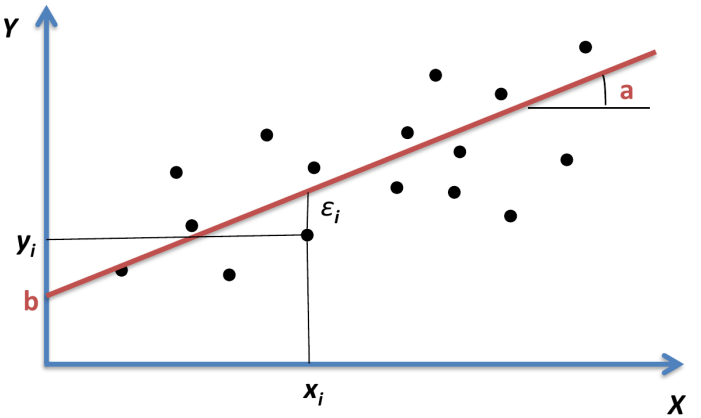
Linear model: general equation

Math equation:

$\hat{Y} = X\beta$

OR

$y_i = a \times x_i + b + \epsilon_i$



With:

- Y and y_i the explained variable
- X and x_i the explanatory variable(s)
- β the model parameters :
  - a the slope
  - b the intercept
- ϵ_i the model residuals (error)

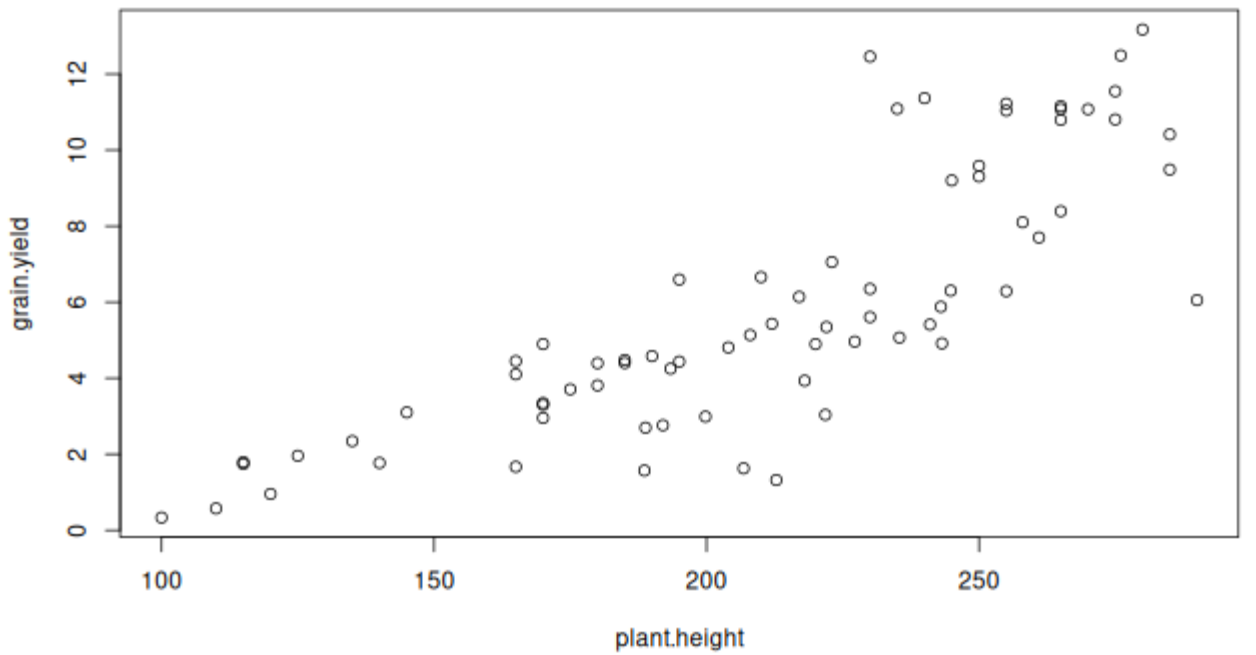
Note:

The regression line is determined by reducing the model residuals

a and b are determined in order to reduce the residual sum of square: $SSres = \sum \epsilon_i^2$

Data representation

```
plot(grain.yield ~ plant.height, data = maize_2012)
```



Using the `lm()` function: simple regression

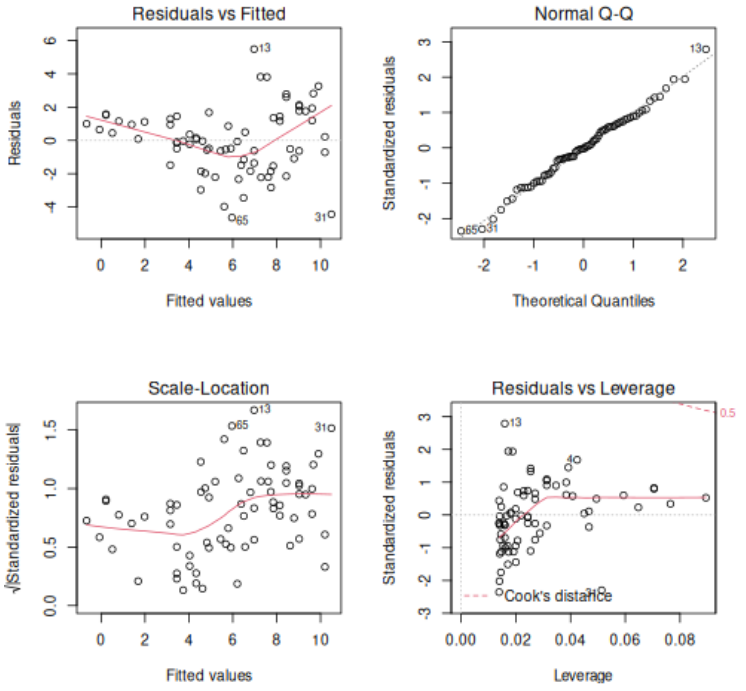
```
mod1 <- lm(grain.yield ~ plant.height, maize_2012)

summary(mod1)
```

```
##
## Call:
## lm(formula = grain.yield ~ plant.height, data = maize_2012)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6392 -1.3942 -0.0377  1.3085  5.4872
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.532477   1.060664  -6.159  4.1e-08 ***
## plant.height  0.058711   0.004892  12.002  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.984 on 70 degrees of freedom
## Multiple R-squared:  0.673,    Adjusted R-squared:  0.6683
## F-statistic:   144 on 1 and 70 DF,  p-value: < 2.2e-16
```

Check model assumptions

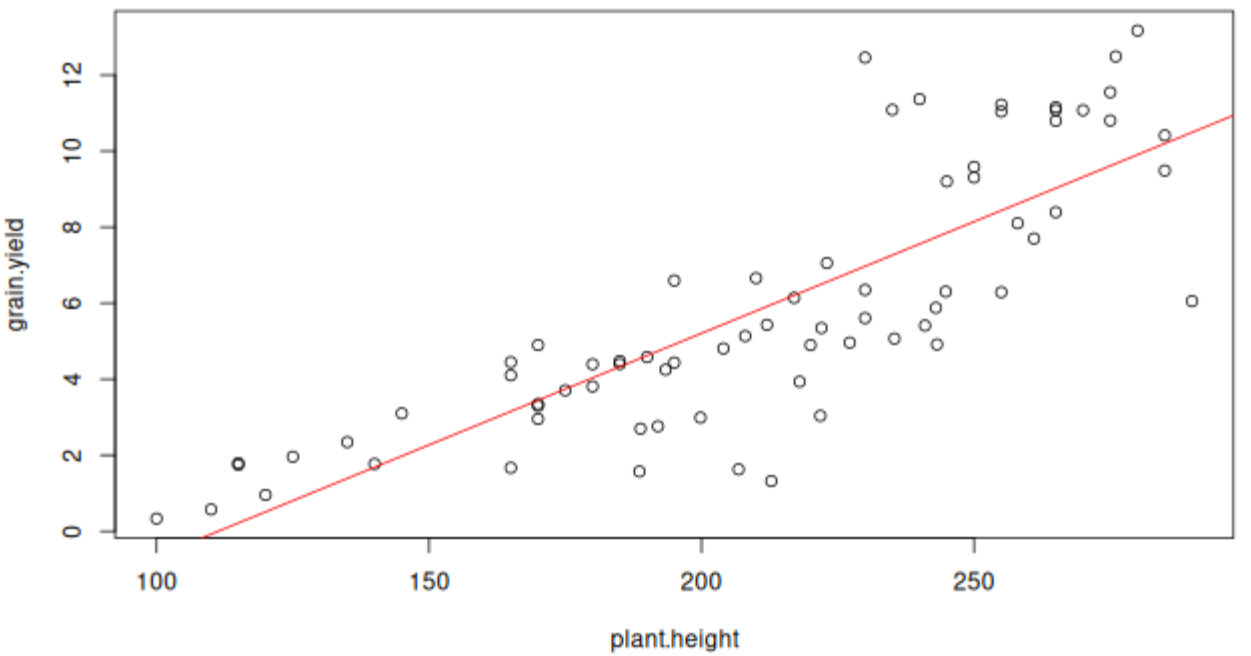
```
par(mfrow = c(2, 2)) ; plot(mod1)
```



Normality test: `shapiro.test` (limit of the normality of residues in linear models here)

Model representation

```
plot(grain.yield ~ plant.height, data = maize_2012)
abline(mod1, col = "red")
```



Model parameters and predictions

Model coefficients:

```
coef(mod1)
```

```
##  (Intercept) plant.height
##  -6.53247740   0.05871075
```

Model parameters and predictions

Model coefficients:

```
coef(mod1)
```

```
##  (Intercept) plant.height
##  -6.53247740   0.05871075
```

Model confidence interval:

```
confint(mod1)
```

```
##                    2.5 %      97.5 %
## (Intercept)  -8.64790461 -4.41705020
## plant.height  0.04895452  0.06846698
```

Model parameters and predictions

Model coefficients:

```
coef(mod1)
```

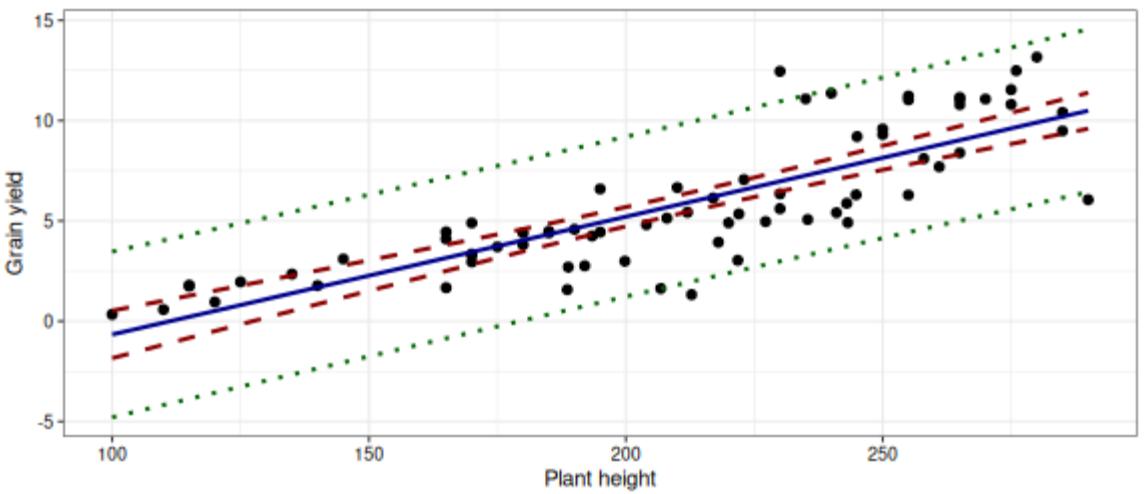
```
##  (Intercept) plant.height
##  -6.53247740   0.05871075
```

Model confidence interval:

```
confint(mod1)
```

```
##                    2.5 %      97.5 %
## (Intercept)  -8.64790461 -4.41705020
## plant.height  0.04895452  0.06846698
```

Model Predictions: `predict(mod1)` with `interval = ""` as argument.



Question 3: What is the effect of the plant height and tassel height on the grain number?

Question 3: What is the effect of the plant height and silking on the grain yield?

Key parts to this question:

- Effect of **TWO numerical** variable on **one numerical** variable
- What are the units?

`plant.height`: plant height from ground to the base of the flag leaf (cm)

`grain.yield`: grain yield adjusted at 15% humidity (t ha^{-1})

`silking`: cumulated thermal time (d20°C) from emergence to flowering

They are all **continuous** variables

Using the `lm()` function: multiple regression

Multiple regression are written as follow:

```
mod2 <- lm(grain.yield ~ plant.height*silking, maize_2012)
```

Note: A $\times$ B is the same as A+B+A:B; **A** and **B**: the main predictors; **A:B** the interaction

Using the `lm()` function: multiple regression

Multiple regression are written as follow:

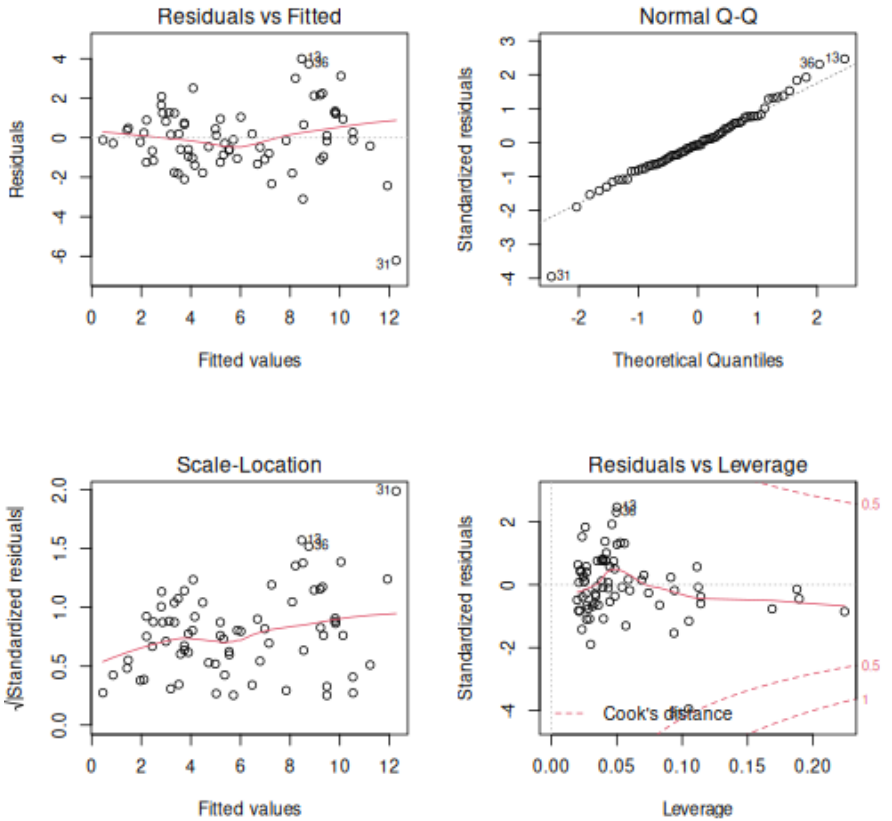
```
mod2 <- lm(grain.yield ~ plant.height*silking, maize_2012)
```

Note: A $\times$ B is the same as A+B+A:B; **A** and **B**: the main predictors; **A:B** the interaction

```
##
## Call:
## lm(formula = grain.yield ~ plant.height * silking, data = maize_2012)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2143 -0.9700 -0.1194  0.9457  3.9959
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -4.068e+01  1.271e+01  -3.201 0.002084 **
## plant.height          2.888e-01  6.053e-02   4.772 1.01e-05 ***
## silking               5.060e-01  1.738e-01   2.912 0.004850 **
## plant.height:silking -3.388e-03  8.462e-04  -4.004 0.000157 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.663 on 68 degrees of freedom
## Multiple R-squared:  0.7767,    Adjusted R-squared:  0.7669
## F-statistic: 78.84 on 3 and 68 DF,  p-value: < 2.2e-16
```

Check model assumptions

```
par(mfrow = c(2, 2)) ; plot(mod2)
```



On problem: (multi) collinearity

When the predictors are **correlated**.

On problem: (multi) collinearity

When the predictors are **correlated**.

Effect:

- Increase estimates standard error
- Lose of prediction power
- Due to confounding variable
  (influence both explicative and explained variables)

On problem: (multi) collinearity

When the predictors are **correlated**.

Effect:

- Increase estimates standard error
- Lose of prediction power
- Due to confounding variable
  (influence both explicative and explained variables)

Problem resolution:

- We don't care if it's for prediction
- Obtain more data
- Eliminate one variable
- Scale the explicatives variables: `mean = 0` and `sd = 1`

Question 4: What is the effect of the water treatment on plant height?

Question 4: What is the effect of the maize variety on plant height?

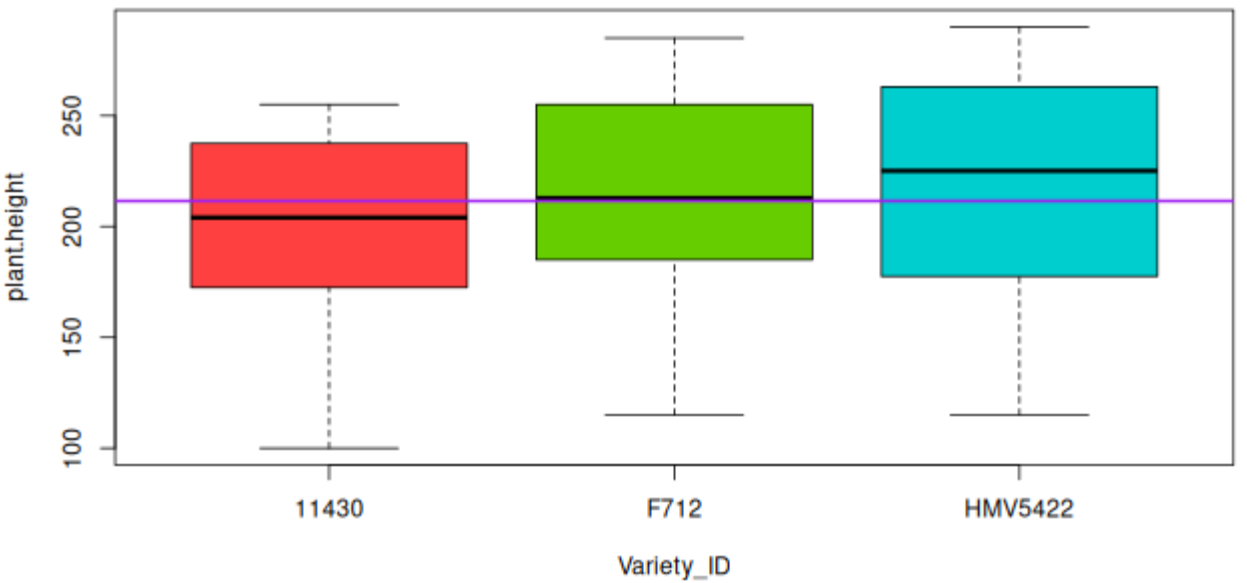
Key parts to this question:

- Effect of **ONE factor** (water regime) on a **numerical** variable (plant height)
- `plant.height` : from ground level to the base of the flag leaf (highest) leaf (cm), continuous variable

Linear model: one factor

Model representation:

```
boxplot(plant.height ~ Variety_ID, maize_2012, col = c("brown1", "chartreuse3", "c
abline(a = mean(maize_2012$plant.height), b = 0, lwd = 2, col = "purple")
```



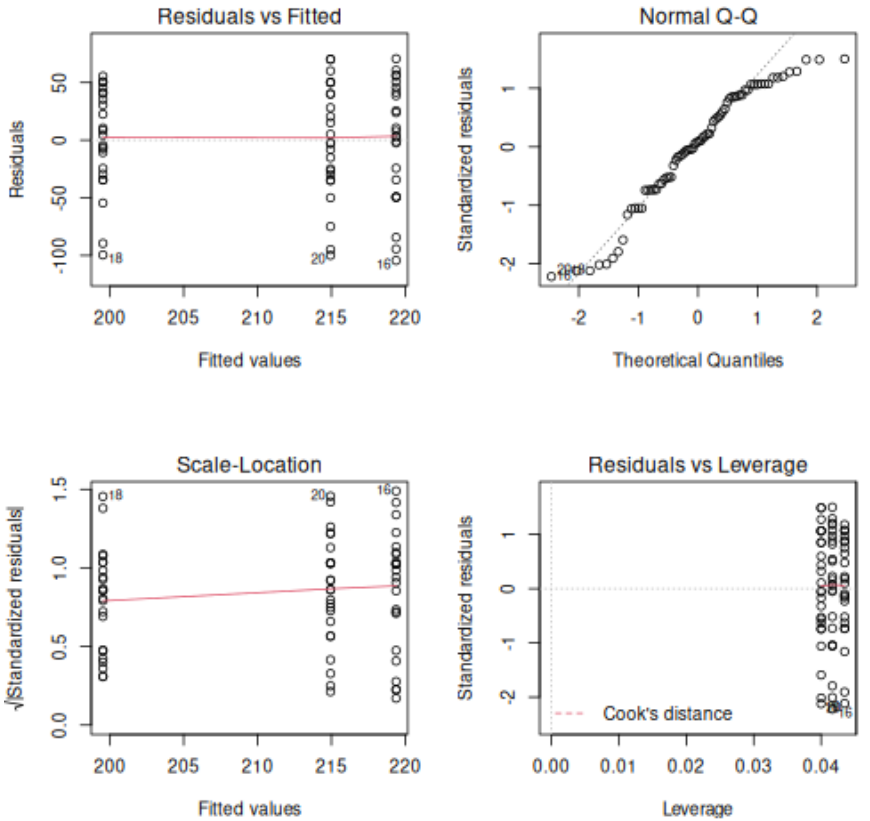
Using the `lm()` function: one factor

```
mod3 <- lm(plant.height ~ Variety_ID, maize_2012)
summary(mod3)
```

```
##
## Call:
## lm(formula = plant.height ~ Variety_ID, data = maize_2012)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.375  -31.046    4.047   40.720   70.625
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         199.53      10.02  19.908   <2e-16 ***
## Variety_IDF712       15.41      13.89   1.109    0.271
## Variety_IDHMV5422    19.84      14.03   1.415    0.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.07 on 69 degrees of freedom
## Multiple R-squared:  0.03088,    Adjusted R-squared:  0.002785
## F-statistic: 1.099 on 2 and 69 DF,  p-value: 0.3389
```

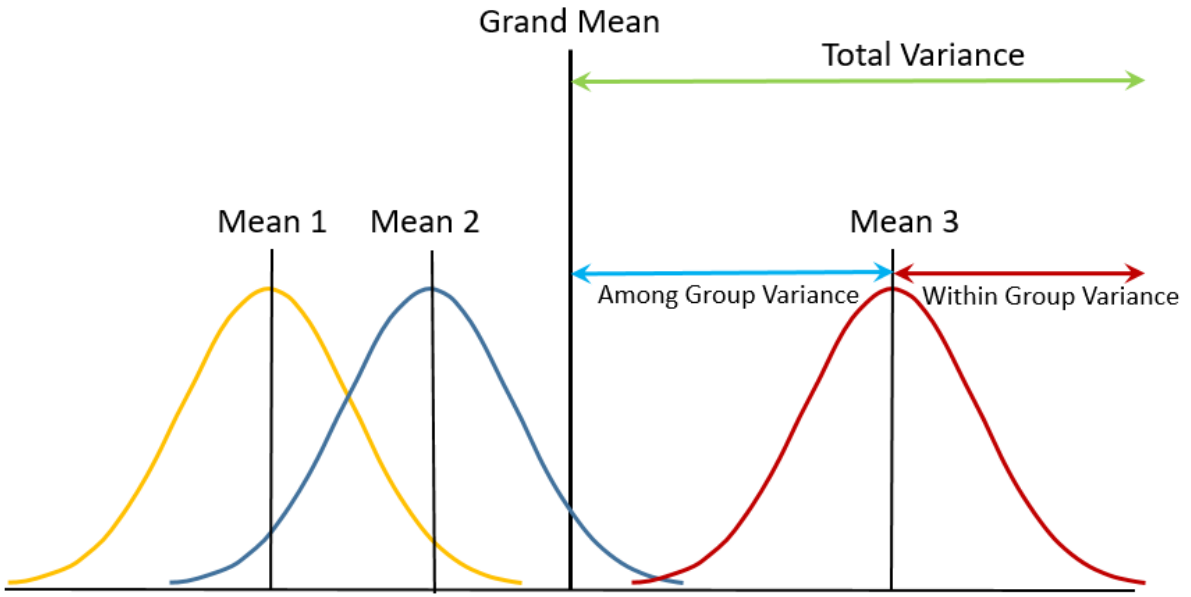
Check model assumptions

```
par(mfrow = c(2, 2)) ; plot(mod3)
```



Analysis of variance (ANOVA)

- Variance can be used to test whether means of a factor differed: as the means become more different, the variances increase



ANOVA Several options:

- `aov` function: (sensitive to unbalanced dataset)

  ```
  mod_aov <- aov(x ~ a)
  ```
  `mod_aov` output is directly the Analysis of Variance Table

- `anova` function:

  ```
  mod_lm <- lm(x ~ a)
  ```
  `anova(mod_lm)` gives the Analysis of Variance Table

Check for balanced data:

Hint: use the `table()` function for our factors to see if we have a balanced dataset for the factor in our dataset.

```
table(maize_2012$Variety_ID)
```

```
##
##   11430    F712 HMV5422
##      23      25      24
```

Note: linear model work with unbalanced date, but effects are more tricky to analysed.

One-way ANOVA

With one factor, the ANOVA detect if the factor within group variance is in the same magnitude as the residual variance (residuals are model error).

We will use the `anova` function: `anova(mod3)`

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variety_ID	2	5.08e+03	2.54e+03	1.1	0.339
Residuals	69	1.59e+05	2.31e+03		

One-way ANOVA

With one factor, the ANOVA detect if the factor within group variance is in the same magnitude as the residual variance (residuals are model error).

We will use the `anova` function: `anova(mod3)`

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variety_ID	2	5.08e+03	2.54e+03	1.1	0.339
Residuals	69	1.59e+05	2.31e+03		

We will decipher the Analysis of Variance table in detail with the next example.

Note 1: an anova is basically an extended t-test
(which compares the variance of two groups)

Note 2: an one-way ANOVA is used when the factor present three or more levels

Order of factor levels

Is the order of factor levels important?

- **YES**: example of fertilization treatment with `low`, `medium`, `high`
- **NO**: example of the flower fragrance with `frangrance 1`, `frangrance 2`

Order of factor levels

Is the order of factor levels important?

- **YES**: example of fertilization treatment with `low`, `medium`, `high`
- **NO**: example of the flower fragrance with `frangrance 1`, `frangrance 2`

In `lm()` function: the `contrast` argument:

Contrast argument	calculation
`contr.treatment()` (default)	compares each factor level to the reference level
`contr.sum()`	compares each factor level to general mean
`contr.poly()`	for one variable, detect polynomial effect
`contre.helmert()`	compares each factor level to the average of the next levels

Order of factor levels

Is the order of factor levels important?

- **YES**: example of fertilization treatment with `low`, `medium`, `high`
- **NO**: example of the flower fragrance with `frangrance 1`, `frangrance 2`

In `lm()` function: the `contrast` argument:

Contrast argument	calculation
`contr.treatment()` (default)	compares each factor level to the reference level
`contr.sum()`	compares each factor level to general mean
`contr.poly()`	for one variable, detect polynomial effect
`contre.helmert()`	compares each factor level to the average of the next levels

Change factor level reference with `relevel` OR `fct_relevel`.

Question 5: What is the effect of the water treatment and the maize variety on grain weight?

Question 5: What is the effect of the water treatment and the varity ID on grain weight?

Key parts to this question:

- Effect of **TWO factors** (water regime and maize variety) on a **numerical** variable (grain weight)
- `grain.weight` : mass of individual grain (g), continuous variable

ANOVA reminder: `x ~ a+b+a:b`

Factor	Type 1 ANOVA	Type 2 ANOVA	Type 3 ANOVA
Factor A	SS(A)	SS(A I B)	SS(A I B, AB)
Factor B	SS(B I A)	SS(B I A)	SS(B I A, AB)
Interaction AB	SS(AB I B, A)	NA	SS(AB I B, A)
Notes	SS(A) + SS(B I A) = SS(A, B)	SS(A I B) + SS(B I A) ≠ SS(A, B)	NA

ANOVA reminder: `x ~ a+b+a:b`

Factor	Type 1 ANOVA	Type 2 ANOVA	Type 3 ANOVA
Factor A	SS(A)	SS(A I B)	SS(A I B, AB)
Factor B	SS(B I A)	SS(B I A)	SS(B I A, AB)
Interaction AB	SS(AB I B, A)	NA	SS(AB I B, A)
Notes	SS(A) + SS(B I A) = SS(A, B)	SS(A I B) + SS(B I A) ≠ SS(A, B)	NA

Type 1:

- H_0 for A: Means for each levels weighted by sample size are equal
- SS depend on factor order
- Sensitive to model balance

Type 2:

- We assume no interaction between A and B

ANOVA reminder: `x ~ a+b+a:b`

Factor	Type 1 ANOVA	Type 2 ANOVA	Type 3 ANOVA
Factor A	SS(A)	SS(A I B)	SS(A I B, AB)
Factor B	SS(B I A)	SS(B I A)	SS(B I A, AB)
Interaction AB	SS(AB I B, A)	NA	SS(AB I B, A)
Notes	SS(A) + SS(B I A) = SS(A, B)	SS(A I B) + SS(B I A) ≠ SS(A, B)	NA

Type 1:

- H_0 for A: Means for each levels weighted by sample size are equal
- SS depend on factor order
- Sensitive to model balance

Type 2:

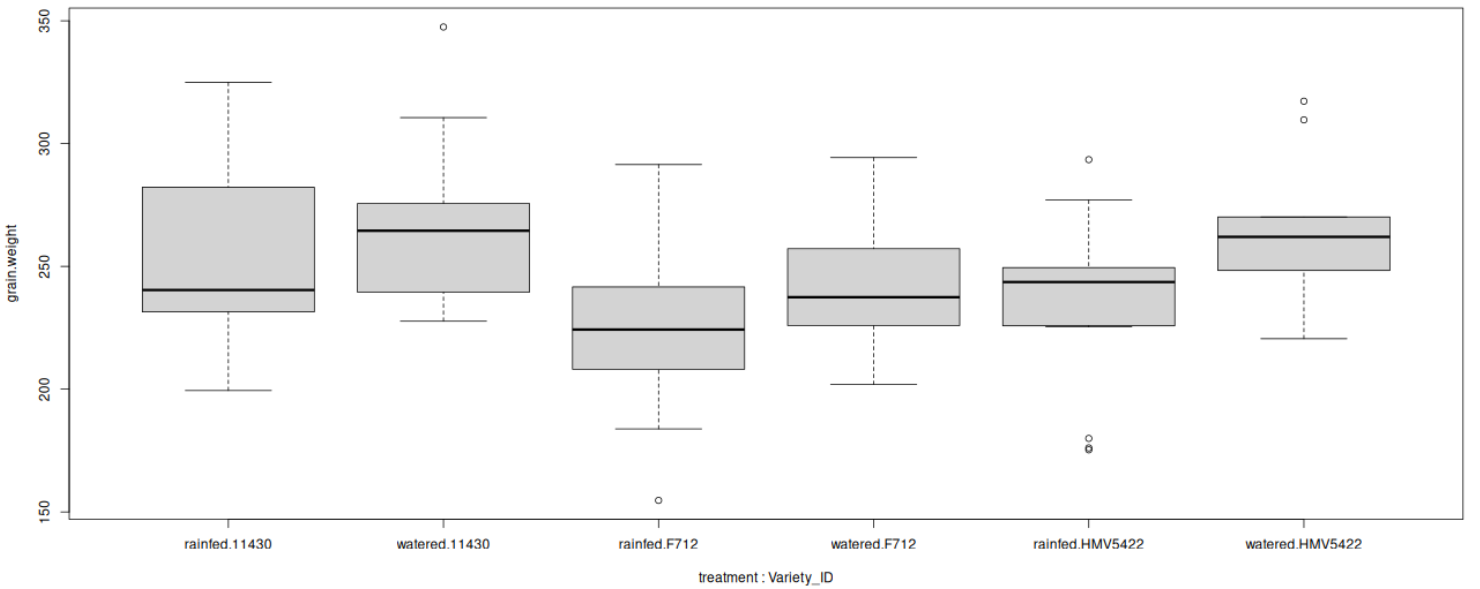
- We assume no interaction between A and B

Type 3:

- For A and B: Means for each levels **NOT** weighted by sample size are equal
- Use the contrast `contr.sum`

Linear model: two factors

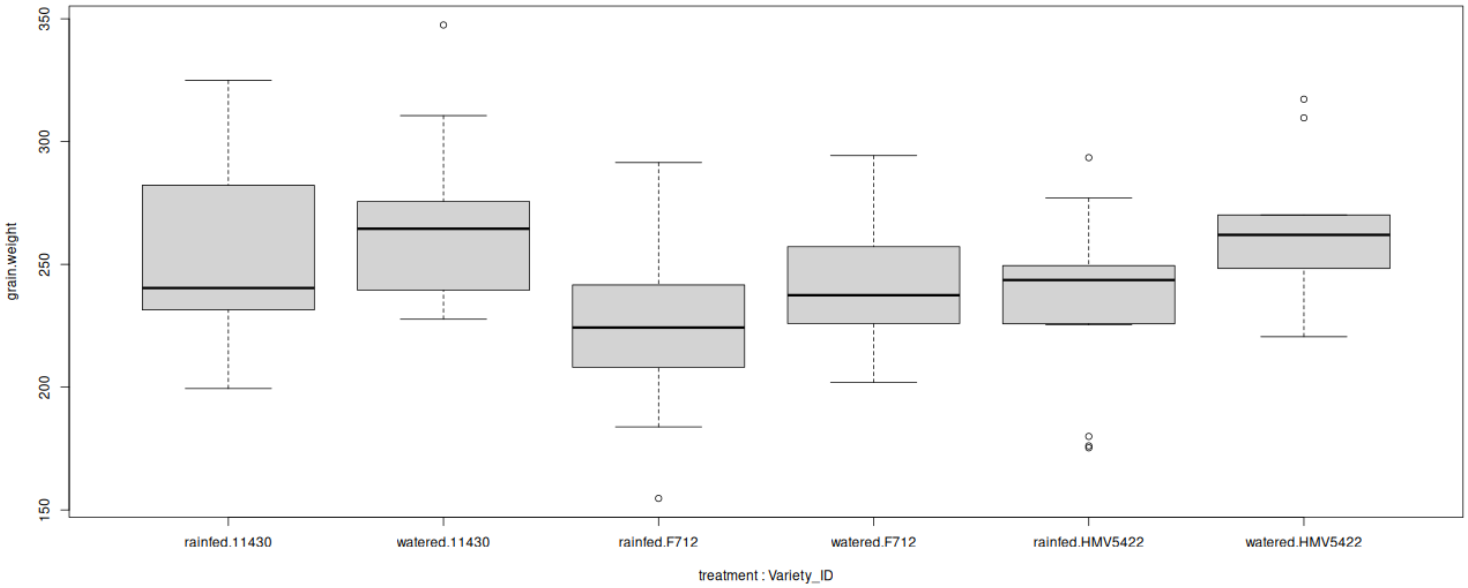
```
boxplot(grain.weight ~ treatment*Variety_ID, data = maize_2012)
```



What can we see?

Linear model: two factors

```
boxplot(grain.weight ~ treatment*Variety_ID, data = maize_2012)
```



What can we see?

- There seems to be a positive effect of watering plants on grain weight depending on maize variety (makes sense). We don't see a clear treatment effect, but there might be.

Testing models

Can we use `aov()`? **Hint:** The `aov` function is a type 1 ANOVA.

Testing models

Can we use `aov()`? **Hint:** The `aov` function is a type 1 ANOVA.

```
replications(grain.weight ~ treatment+Variety_ID, data = maize_2012)
```

```
## $treatment
## treatment
## rainfed watered
##      44      28
##
## $Variety_ID
## Variety_ID
##   11430    F712 HMV5422
##      23      25      24
```

Because we have unbalanced data, it is better to use the `lm()` function.

Testing models

Can we use `aov()`? **Hint:** The `aov` function is a type 1 ANOVA.

```
replications(grain.weight ~ treatment+Variety_ID, data = maize_2012)
```

```
## $treatment
## treatment
## rainfed watered
##      44      28
##
## $Variety_ID
## Variety_ID
##   11430    F712 HMV5422
##      23      25      24
```

Because we have unbalanced data, it is better to use the `lm()` function.

Notes: for type 2 and 3 ANOVA, use the `Anova` function from the `car` package.

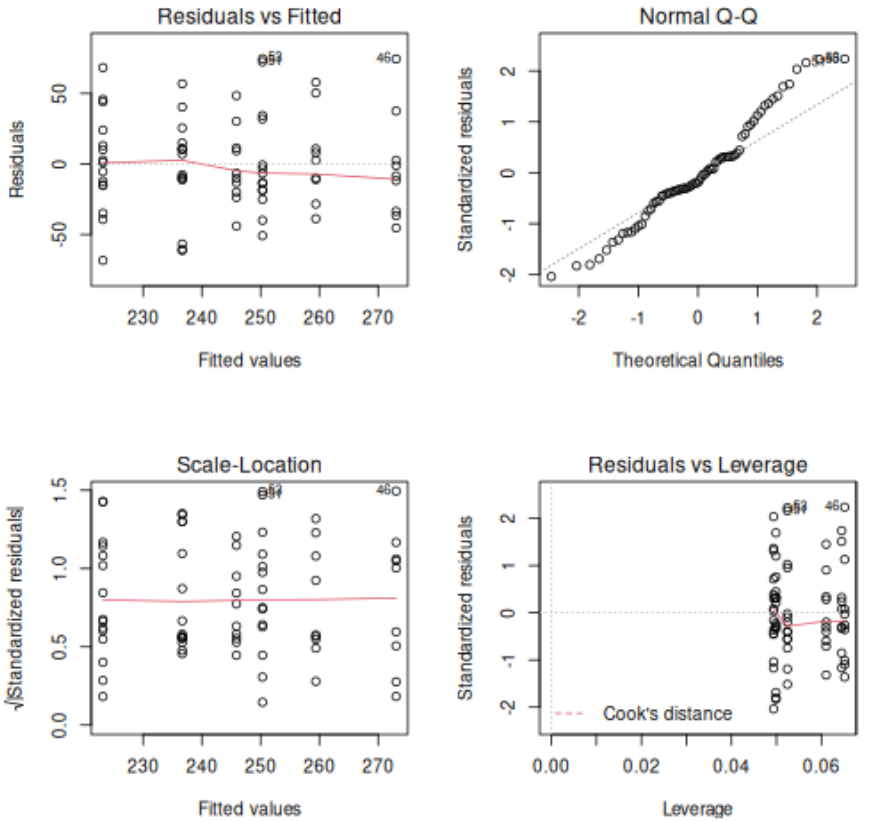
Using the `lm()` function: two factor

```
mod4 <- lm(grain.weight ~ treatment + Variety_ID, data = maize_2012)
summary(mod4)
```

```
##
## Call:
## lm(formula = grain.weight ~ treatment + Variety_ID, data = maize_2012)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.444 -18.533  -6.029  13.452  74.599
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        250.316      7.889  31.731  < 2e-16 ***
## treatmentwatered    22.735      8.331   2.729  0.00808 **
## Variety_IDF712     -27.158      9.954  -2.728  0.00810 **
## Variety_IDHMV5422  -13.660     10.054  -1.359  0.17875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.45 on 68 degrees of freedom
## Multiple R-squared:  0.1786,    Adjusted R-squared:  0.1423
## F-statistic: 4.928 on 3 and 68 DF,  p-value: 0.003717
```

Check the assumptions

```
par(mfrow = c(2, 2)) ; plot(mod4)
```



Looking at the analysis of variance table

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	1	8.71e+03	8.71e+03	7.34	0.00854
Variety_ID	2	8.84e+03	4.42e+03	3.72	0.0292
Residuals	68	8.07e+04	1.19e+03		

Looking at the analysis of variance table

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	1	8.71e+03	8.71e+03	7.34	0.00854
Variety_ID	2	8.84e+03	4.42e+03	3.72	0.0292
Residuals	68	8.07e+04	1.19e+03		

Theoretical calculation:

Source	SS	df	MS	F
Effect A	given	a-1	SS/df	MS(A) / MS(W)
Effect B	given	b-1	SS/df	MS(B) / MS(W)
Effect A:B	given	(a-1)(b-1)	SS/df	MS(A:B) / MS(W)
Within (residuals)	given		SS/df	

Let's break down the table output

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	1	8.71e+03	8.71e+03	7.34	0.00854
Variety_ID	2	8.84e+03	4.42e+03	3.72	0.0292
Residuals	68	8.07e+04	1.19e+03		

Df: degrees of freedom = the number of values in the final calculation of a statistic that are free to vary

- Each main effect has k-1 Df

ex: `Variety_ID` has 3 levels (k=3) so df = 3-1 = 2

and `treatment` has two levels (m=2) so df = 2-1 = 1

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	1	8.71e+03	8.71e+03	7.34	0.00854
Variety_ID	2	8.84e+03	4.42e+03	3.72	0.0292
Residuals	68	8.07e+04	1.19e+03		

- The residuals df is the total number of observations (N) - 1 - the df of each effect (k-1) and (m-1).
- You can find N by calculating the `length` of `grain.weight` in the maize_2012 data table, or by looking in the Global Environment data values

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	1	8.71e+03	8.71e+03	7.34	0.00854
Variety_ID	2	8.84e+03	4.42e+03	3.72	0.0292
Residuals	68	8.07e+04	1.19e+03		

- The residuals df is the total number of observations (N) - 1 - the df of each effect (k-1) and (m-1).
- You can find N by calculating the `length` of `grain.weight` in the maize_2012 data table, or by looking in the Global Environment data values

```
length(maize_2012$grain.weight)
```

```
## [1] 72
```

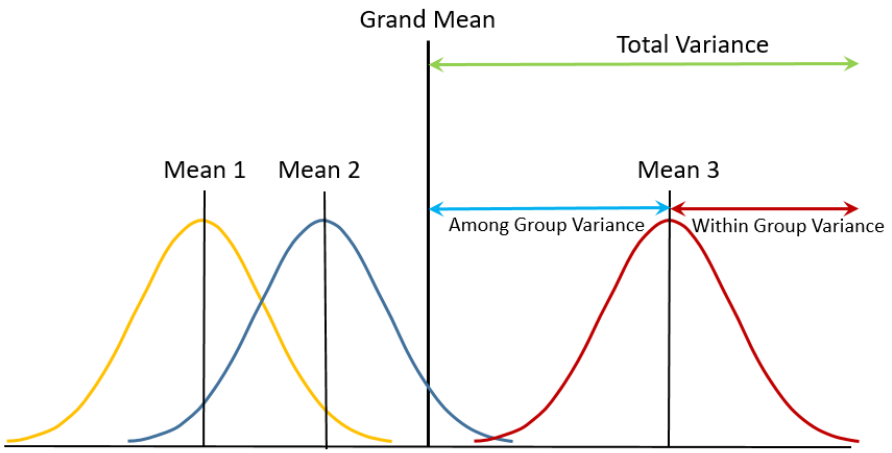
= N - 1 - (k-1) - (m-1) = 72 - 1 - 2 - 1 = 68

Why is this important? You can use this to better understand tables published in papers. You can also easily see if authors have removed data points from their analysis (the N.)

Fisher's test (Fisher-Snedecor)

Sum of Squares difference between variables / Sum of squares residuals (within)

Ex: If the `Variety_ID` means of grain weight are more different (vary more between the different levels) than the grain weight vary within the different Variety levels, the F-test ratio would be greater than an F-value expected by chance.



The p-value

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	1	8.71e+03	8.71e+03	7.34	0.00854
Variety_ID	2	8.84e+03	4.42e+03	3.72	0.0292
Residuals	68	8.07e+04	1.19e+03		

R then compares the F ratio calculated to the expected value given the two degrees of freedom (between the groups and within the groups), with alpha set to 0.05%.

If the p-value is less than the chosen significance level, the data provide sufficient evidence to conclude that your model fits the data better than a model with no independent variables.

Be cautious with interpreting p-values (will go over next week)

Post-hoc tests

The F-test doesn't reveal which Variety_ID means are different, only that the Variety_ID means are different.

Therefore, we can do a multiple comparison test, also called a posthoc test

Post-hoc tests

The F-test doesn't reveal which Variety_ID means are different, only that the Variety_ID means are different.

Therefore, we can do a multiple comparison test, also called a posthoc test

```
modposthoc = lm(grain.weight ~ Variety_ID+treatment,
                data = maize_2012, na.action=na.omit)
tuk <- glht(modposthoc , linfct = mcp(Variety_ID= "Tukey"))
```

```
summary(tuk)
```

```
##
##      Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = grain.weight ~ Variety_ID + treatment, data = maize_2012,
##     na.action = na.omit)
##
## Linear Hypotheses:
##                       Estimate Std. Error t value Pr(>|t|)
## F712 - 11430 == 0     -27.158      9.954  -2.728   0.0218 *
## HMV5422 - 11430 == 0  -13.660     10.054  -1.359   0.3682
## HMV5422 - F712 == 0    13.498      9.848   1.371   0.3619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
summary(tuk)
```

```
##
##      Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = grain.weight ~ Variety_ID + treatment, data = maize_2012,
##     na.action = na.omit)
##
## Linear Hypotheses:
##                       Estimate Std. Error t value Pr(>|t|)
## F712 - 11430 == 0      -27.158      9.954  -2.728   0.0218 *
## HMV5422 - 11430 == 0   -13.660     10.054  -1.359   0.3682
## HMV5422 - F712 == 0     13.498      9.848   1.371   0.3619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We can also use the `cld()` function to get the classic letter display.

```
tuk.cld <- cld(tuk)   # letter-based display
tuk.cld
```

```
##   11430    F712 HMV5422
##     "b"     "a"    "ab"
```

```
summary(tuk)
```

```
##
##      Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = grain.weight ~ Variety_ID + treatment, data = maize_2012,
##     na.action = na.omit)
##
## Linear Hypotheses:
##                       Estimate Std. Error t value Pr(>|t|)
## F712 - 11430 == 0      -27.158      9.954  -2.728   0.0218 *
## HMV5422 - 11430 == 0   -13.660     10.054  -1.359   0.3682
## HMV5422 - F712 == 0     13.498      9.848   1.371   0.3619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

We can also use the `cld()` function to get the classic letter display.

```
tuk.cld <- cld(tuk)   # letter-based display
tuk.cld
```

```
##   11430    F712 HMV5422
##     "b"     "a"    "ab"
```

Note: for A:B interaction Tukey test, look at the `emmeans` package and function.

New dataset for better understanding the ANCOVA

The Ipomopsis data:

```
Ipo <- read_csv2("Data/Exemple/ipomopsis.csv") %>%
  mutate(Grazing = as_factor(Grazing),
         Grazing = fct_relevel(Grazing, c("Grazed", "Ungrazed")))
```

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.

## Rows: 40 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ";"
## chr (1): Grazing
## dbl (2): Root, Fruit
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Ipomopsis: North-america flower genus, mountainous ecosystem (sub-alpine Rocky mountains)

Question 6: What are the effect of the grazing treatment and roots length on fruit production?

Question 6: What are the effect of the grazing treatment and roots length on fruit production?

Key parts to this question:

- Effect of **ONE factor** (grazing) and **ONE continuous** variable (number of fruits) on a **numerical** variable (root length)
- `Root`: root size (cm), continuous variable
- `Fruit`: average number of fruits (n), continuous variable
- `Grazing`: grazed or ungrazed plant (factor), discrete variable

Question 6: What are the effect of the grazing treatment and roots length on fruit production?

Key parts to this question:

- Effect of **ONE factor** (grazing) and **ONE continuous** variable (number of fruits) on a **numerical** variable (root length)
- `Root`: root size (cm), continuous variable
- `Fruit`: average number of fruits (n), continuous variable
- `Grazing`: grazed or ungrazed plant (factor), discrete variable

Which test do you think we should use?

Question 6: What are the effect of the grazing treatment and roots length on fruit production?

Key parts to this question:

- Effect of **ONE factor** (grazing) and **ONE continuous** variable (number of fruits) on a **numerical** variable (root length)
- `Root`: root size (cm), continuous variable
- `Fruit`: average number of fruits (n), continuous variable
- `Grazing`: grazed or ungrazed plant (factor), discrete variable

Which test do you think we should use?

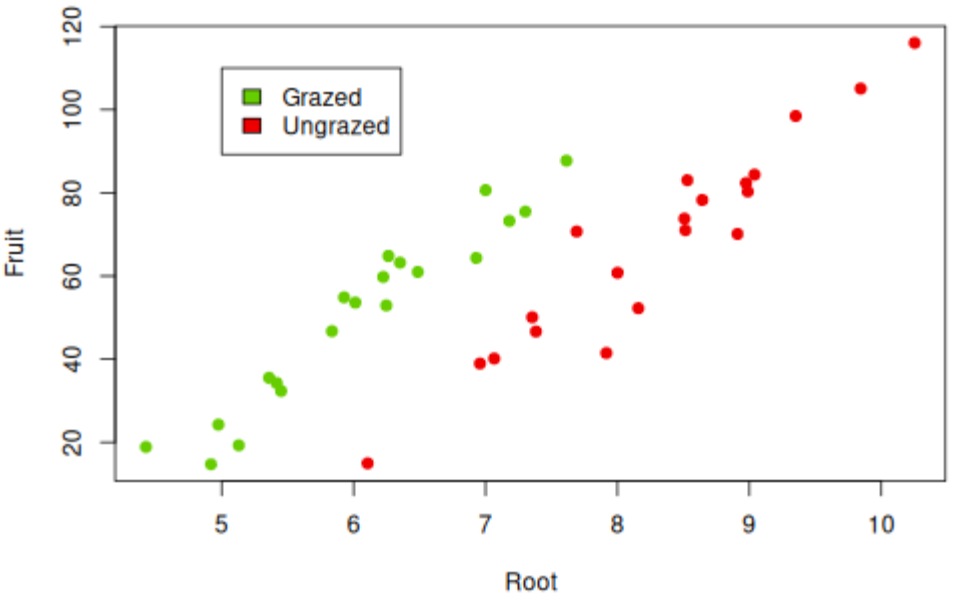
Analysis of covariance (ANCOVA):

An ANCOVA is the combination of a linear regression and an ANOVA.

Look at the data

```
table(Ipo$Grazing)
```

```
col1 <- ifelse(Ipo$Grazing == "Grazed", "red2", "chartreuse3")
plot(Fruit ~ Root, Ipo, col = col1, pch = 19)
legend(5, 110, c("Grazed", "Ungrazed"), fill = c("chartreuse3", "red2"))
```



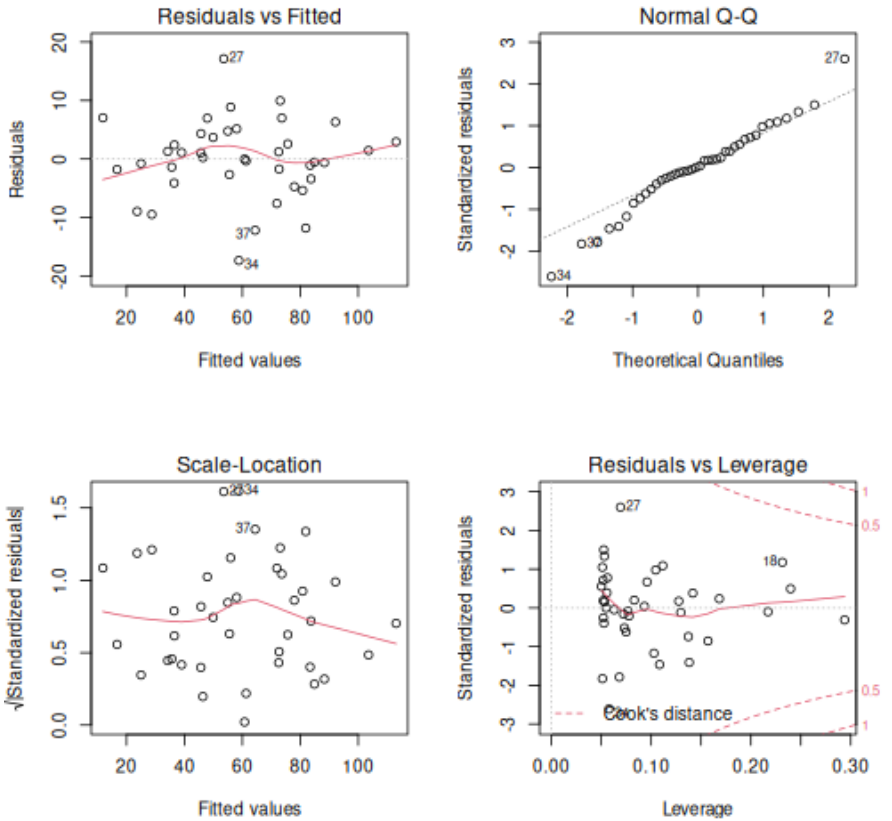
Using the `lm()` function: two factor

```
mod5 <- lm(Fruit ~ Root * Grazing, Ipo)
summary(mod5)
```

```
##
## Call:
## lm(formula = Fruit ~ Root * Grazing, data = Ipo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3177  -2.8320   0.1247   3.8511  17.1313
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -125.173     12.811  -9.771 1.15e-11 ***
## Root                   23.240      1.531  15.182  < 2e-16 ***
## GrazingUngrazed        30.806     16.842   1.829   0.0757 .
## Root:GrazingUngrazed    0.756      2.354   0.321   0.7500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.831 on 36 degrees of freedom
## Multiple R-squared:  0.9293,    Adjusted R-squared:  0.9234
## F-statistic: 157.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

Check the assumptions

```
par(mfrow = c(2, 2)) ; plot(mod5)
```



Performed the ANCOVA

```
anova(mod5)
```

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Root	1	1.68e+04	1.68e+04	360	2.49e-20
Grazing	1	5.26e+03	5.26e+03	113	1.21e-12
Root:Grazing	1	4.81	4.81	0.103	0.75
Residuals	36	1.68e+03	46.7		

Performed the ANCOVA

```
anova(mod5)
```

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Root	1	1.68e+04	1.68e+04	360	2.49e-20
Grazing	1	5.26e+03	5.26e+03	113	1.21e-12
Root:Grazing	1	4.81	4.81	0.103	0.75
Residuals	36	1.68e+03	46.7		

What are the conclusion?

Performed the ANCOVA

```
anova(mod5)
```

rownames	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Root	1	1.68e+04	1.68e+04	360	2.49e-20
Grazing	1	5.26e+03	5.26e+03	113	1.21e-12
Root:Grazing	1	4.81	4.81	0.103	0.75
Residuals	36	1.68e+03	46.7		

What are the conclusion?

- Less fruits are produced when the plants are grazed
- Grazing do not influence the fruit production, increasing with root size

Model selection

Build a second model without interaction:

```
mod5.b <- lm(Fruit ~ Root + Grazing, Ipo)
```

Model selection

Build a second model without interaction:

```
mod5.b <- lm(Fruit ~ Root + Grazing, Ipo)
```

You can check the model summary and assumptions.

Model selection

Build a second model without interaction:

```
mod5.b <- lm(Fruit ~ Root + Grazing, Ipo)
```

You can check the model summary and assumptions.

Which model return the best description of the data? `mod5` or `mod5.b`?

`anova(mod5, mod5.b)`

rownames	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	36	1.68e+03				
2	37	1.68e+03	-1	-4.81	0.103	0.75

Model selection

Build a second model without interaction:

```
mod5.b <- lm(Fruit ~ Root + Grazing, Ipo)
```

You can check the model summary and assumptions.

Which model return the best description of the data? `mod5` or `mod5.b`?

`anova(mod5, mod5.b)`

rownames	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	36	1.68e+03				
2	37	1.68e+03	-1	-4.81	0.103	0.75

None is the best. Always select the most sparing solution.

Model selection

Build a second model without interaction:

```
mod5.b <- lm(Fruit ~ Root + Grazing, Ipo)
```

You can check the model summary and assumptions.

Which model return the best description of the data? `mod5` or `mod5.b`?

`anova(mod5, mod5.b)`

rownames	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	36	1.68e+03				
2	37	1.68e+03	-1	-4.81	0.103	0.75

None is the best. Always select the most sparing solution.

The `mod5.b` should be used.

Your turn

In groups of 2:

- State a research question for this data set (a response variable, and then one or more explicative variable)
- Divide the original data set by site, year, treatment and/or variety
- Choose the (or several) fixed effect factors and the response variable according to your research question
- Write your model, test the assumptions and state your results
- Calculate the n, mean, and standard error for your groups

Next week

- Generalized linear models
- Pseudoreplication
- Heritability
- (Introduction to non-linear models and generalized additive models)

References

Schumaker & Tomek 2013. Understanding Statistics Using R. *Springer*

http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram

https://howecoresearch.blogspot.com/2019/01/using-analysis-of-variance-anova-in.html

Ressources:

https://www.youtube.com/watch?v=-yQb_ZJnFXw